Harmonizing the Scale: An End-to-End Self-Supervised Approach for Cross-Modal Data Retrieval in Histopathology Archives

by

Danial Maleki

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2023

© Danial Maleki 2023

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Jungong Han Professor, Department of Computer Science, University of Sheffield, UK
Supervisor(s):	Hamid Tizhoosh Professor, Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA
	Shahryar Rahnamayan Department of Engineering, Brock University
Internal Member:	Nima Maftoon Assistant Professor, Department of Systems Design Engineering, University of Waterloo
	Parsin Haji Reza Professor, Department of Systems Design Engineering, University of Waterloo
Internal-External Member:	Mark Crowley Associate Professor, Department of Electrical and Computer Engineering, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In recent years, the exponential growth of data across various domains has necessitated the development of advanced techniques to process and analyze multi-modal big data. This is particularly relevant in the medical domain where data comes in diverse formats, such as images, reports, and molecular data. Consequently, *bidirectional cross-modal data retrieval* has become crucial for numerous research disciplines and domains.

Cross-modal retrieval seeks to identify a shared latent space where different modalities, such as image-text pairs, are closely related. Obtaining high-quality vision and text embeddings is vital for achieving this objective. Although training language models is feasible due to the availability of public data and the absence of labelling requirements, training vision models to generate effective embeddings can be challenging due to the scarcity of labelled data when relying on supervised models. To address this challenge, an end-to-end approach to learning vision embeddings in a self-supervised manner, coined *H-DINO+LILE*, is introduced through a modification of the DINO model.

The suggested innovation to improve the DINO model involves transforming the existing local and global patching scheme into a new *harmonizing* patching approach, termed *H-DINO*, where the magnitude of various augmentations is consistently maintained. This method captures the contextual information of images more consistently, thereby improving feature representation and retrieval accuracy. Furthermore, a unique architecture is proposed that integrates self-supervised learning and cross-modal retrieval modules in a back-to-back configuration, enabling improved representation of cross-modal and individual modalities using self-attention and cross-attention modules. This architecture features end-to-end training with a new loss term that facilitates image and text representation in the joint latent space.

The efficacy of the proposed framework is validated on various private and public datasets across diverse tasks such as patch-based (sub-images) and WSI-based (whole slide images) retrieval, as well as text retrieval tasks. This thesis demonstrates that the proposed framework significantly bolsters cross-modal retrieval within the medical domain. Moreover, its applicability extends beyond the medical field, as it can be utilized in other domains that require cross-modal retrieval and contain patching of gigapixel images in their methodologies.

Acknowledgements

Throughout the journey of writing this thesis, I have been fortunate to receive abundant support and assistance. First and foremost, I extend my heartfelt gratitude to my supervisors, Professor H.R. Tizhoosh and Shahryar Rahnamayan. Their invaluable expertise played a pivotal role in shaping the research questions and methodology. Their insightful feedback and guidance continuously pushed me to refine my thinking and elevate the quality of my work. Their unwavering encouragement and patience have been a source of inspiration throughout this project.

I am deeply indebted to my colleagues at KIMIA Lab for their exceptional collaboration and camaraderie.

A special acknowledgment goes to Mitra for her unparalleled support. Additionally, I am immensely grateful to my dear friends Soheila, Amir, Mohammad and Milad, who stood by me through the challenging times, cheered me on, and celebrated every accomplishment.

I also express my sincere appreciation to my committee members, Prof. Han, Dr. Crowley and Dr. Maftoon and Prof. Haji Reza, for their valuable guidance during my academic journey. Their mentorship provided me with the necessary tools to navigate in the right direction and successfully complete my dissertation.

Last but not least, I want to thank my parents for their wise counsel and unwavering support. Your presence and understanding have been my rock throughout this academic pursuit.

Dedication

This is dedicated to my parents.

Table of Contents

Examining Committee ii

Author's Declaration iii

Abstract iv

Acknowledgements v

Dedication vi

List of Figures x

List of Tables xv

List of Abbreviations xvi

1 Overview 1

1.1 Motivation 1

1.2 Objective 4

1.3 Thesis Outline 6

2 Background and Related Works 7

2.1 Transformer Models 8

2.1.1 Input Embedding . . . 11
2.1.2 Positional Embedding . . . 12
2.1.3 Encoder . . . 14
2.1.4 Attention . . . 15
2.1.5 Transformer Models in Computer Vision . . . 17
2.2 Self-Supervised Learning (SSL) . . . 18
2.3 Cross-Modality Retrieval . . . 19
2.4 Cross-Modality Retrieval in Histopathology . . . 23

3 Cross-Modal Retrieval **26**
3.1 Introduction . . . 26
3.2 Feature Representation . . . 27
3.2.1 Image Representation . . . 29
3.2.2 Text Representation . . . 29
3.3 Multi-Head Self-Attention . . . 31
3.4 Cross-Attention and Memory Block . . . 33
3.5 Iterative Matching Strategy . . . 35
3.6 Employment of the Triplet-Loss Function . . . 37
3.7 Empirical Evaluation and Result Analysis . . . 38
3.8 Benchmark Datasets and Evaluation Criteria . . . 39
3.8.1 Benchmark Datasets . . . 39
3.8.2 Evaluation Metric . . . 40
3.9 Implementation Details . . . 41
3.9.1 Experiments and Results . . . 43

4 Self-Supervised Learning **47**
4.1 Introduction . . . 47
4.2 Feature Representation . . . 48

4.2.1 Image Representation 49
4.2.2 Patching 51
4.3 Loss Function 54
4.4 Datasets and Experiments 56
4.4.1 Benchmark Datasets 57
4.4.2 Evaluation Metric 59
4.4.3 Experiments 60
4.5 Summary 80

5 Summary and Future Work **82**
5.1 Summary 82
5.2 Future Works 84

References **85**

List of Figures

1.1 Example for image-caption pair. The caption for this image reads: "*A timed exposure of the first Space Shuttle mission, STS-1, at Launch Pad A, Complex 39, turns the space vehicle and support facilities into a night-time fantasy of light. To the left of the Shuttle are the fixed and the rotating service structures*". 4

2.1 An Long-Short Term Memory (LSTM)'s repeating module is organized into four interacting layers. Each line contains a full vector from one node's output to the inputs of others. The pink circles indicate point-wise operations such as vector addition, and the yellow boxes represent learned neural network layers. Lines that merge indicate concatenation while forking denotes their content being copied and the copies being sent to distinct destinations. Figure adopted from [11] 10

2.2 The Transformer architecture. The main components are the encoder and decoder parts. The embedding layer and positional embedding are present in both the encoder and the decoder. Aside from that, the Encoder stack is composed of several encoders. Each encoder has a multi-head attention module as well as a feed-forward layer. The final encoder is fed into the decoder stack, which comprises a number of decoders. Each decoder has two multi-head attention layers as well as a feed-forward layer. Image adapted from [143] 11

2.3 The Embedding layer. The word embedding layer will map each token to a vector spanned in d dimensions (the embedding dimension). n is the number of entries in the dictionary, and d is the vector dimension. 13

2.4 (a): Scaled dot-product attention module. (b): Multi-head attention. The attention module performs its computations in parallel many times. Each of these is referred to as an *attention Head*. The attention module divides its Query, Key, and Value arguments N times and feeds each split via a separate Head. The results of all of these identical attention computations are then added to generate a final attention score. Figures adapted from [143] 16

2.5 The DINO model utilizes different distinct random transformations of an input image (for simplicity, only two of them are shown in this Figure), which are respectively passed to the student and teacher networks. While both networks share the same architecture, their parameters are distinct. The output of the teacher network is centred using a batch-wide mean. Each network generates a K-dimensional feature that is normalized using a temperature softmax over the dimension of the feature. Then, a cross-entropy loss is used to determine their degree of similarity. Stop-gradient (sg) operators are applied to the teacher in order to propagate gradients only through the student. The parameters of the teacher are updated with an exponential moving average (ema) of the parameters of the student model. Figure adopted from [32]. 20

3.1 The architecture of Look In-Depth before Looking Elsewhere (LILE) for image-text data is shown. It comprises feature extractors dedicated to each modality, with a vision transformer for images and a text transformer for texts. These feature extractors are followed by two networks with identical architecture, facilitating the integration of information from both modalities. 28

3.2 Overview of the Vision Transformer (ViT) architecture: The input image is divided into a set of fixed-size patches, and each patch is linearly embedded. The total number of patches is denoted as P. 30

3.3 The Roberta architecture leverages tokenization to break down the input text data into individual tokens and assigns each token its corresponding embedding vector, denoted as E. The resulting embedding vectors are then passed through the encoding part of the model. The encoder is comprised of a stack of Transformer encoder blocks, with each block consisting of a self-attention module followed by a feed-forward layer. It is important to note that in this figure, only two encoder blocks are depicted, representing a subset of the complete encoder stack. 32

3.4 Representation of the N-pairs triplet-loss function. The arrows denote the enforced push and pull actions of the loss function. Pull connections are indicated in green, while push connections are marked in red. 38

3.5 Depiction of the similarity matrix for a mini-batch with a size of n. The element $I_i.T._j$ in the matrix reveals the similarity score between the image I_i and the text T_j. Since the paired data have the same subscript, the similarity matrix is inherently diagonal. 39

3.6 A selection of three diverse samples from the ARCH dataset. These images reflect the dataset's broad range of staining techniques and normalization processes. Alongside each image, its associated description showcases the varied structure and depth of narrative detail that characterizes the dataset. This diversity not only speaks to the robustness of the dataset but also highlights the complexity of the task at hand - aligning visual and textual modalities that span a wide array of contexts and styles. 40

3.7 (a) Frequency distribution of the token lengths in the descriptions of the ARCH dataset, showcasing the variability in text lengths. (b) Frequency distribution of the caption lengths in the MS-COCO dataset, highlighting the range of text lengths. The MS-COCO dataset exhibits an average of 13.81 tokens per caption, with a maximum of 65 tokens, while the ARCH dataset demonstrates an average of 51.49 tokens per description, with a maximum of 484 tokens. These statistics emphasize the substantial differences in text length between the two datasets. 42

4.1 Diagram of the H-DINO Architecture: Illustrating both the student and teacher networks. For the sake of clarity, only one patch per view is represented. 51

4.2 Diagrammatic Demonstration of Patch Extraction and Expansion from a WSI for Augmentation: Progressing from left to right, the first step in the process is depicted - the mask and the WSI are overlaid to highlight the region from which patches are extracted. For ease of understanding, a single extracted patch is displayed. The next phase of the procedure is demonstrated in the subsequent image, where the size of the extracted patch is expanded. This enlarged output is then fed into the final transformation stage. 52

4.3 Visual Representation of the View Selection and Augmentation Pipeline: The illustration on the left depicts the process of random view selection, demonstrating the extraction of two views each for the Hyper augmentation (τ_H) and Mild augmentation (τ_M) sets. For clarity, only two instances of each set are displayed. These views are randomly cropped from an expanded patch region, each measuring 224 by 224 pixels. The diagram on the right portrays the augmentation pipeline, detailing two unique augmentation sets with varying intensity levels chosen for the Hyper augmentation and Mild augmentation pipelines. The impact of these differing intensities is further highlighted by the distinct outputs they produce. 53

4.4 This confusion matrix depicts the retrieval performance of the proposed method for **patch-level** classification across 22 distinct primary diagnoses, specifically at an **R@1** recall level. It provides a comprehensive visual representation of how accurately the model classifies each diagnosis and demonstrates how the results are distributed across true and predicted classifications, hence offering a deeper understanding of the model's performance. . 65

4.5 This confusion matrix depicts the retrieval performance of the proposed method for **patch-level** classification across 22 distinct primary diagnoses, specifically at an **R@3** recall level. It provides a comprehensive visual representation of how accurately the model classifies each diagnosis and demonstrates how the results are distributed across true and predicted classifications, hence offering a deeper understanding of the model's performance. . 66

4.6 This confusion matrix depicts the retrieval performance of the proposed method for **patch-level** classification across 22 distinct primary diagnoses, specifically at an **R@5** recall level. It provides a comprehensive visual representation of how accurately the model classifies each diagnosis and demonstrates how the results are distributed across true and predicted classifications, hence offering a deeper understanding of the model's performance. . 67

4.7 This confusion matrix provides a visualization of the retrieval performance at the **R@1** level of the proposed method for the **WSI-level** classification across 22 distinct primary diagnoses. It serves as a detailed display of the model's accuracy in diagnosis classification. By presenting the distribution of results across true and predicted classes, it enables a profound understanding of the overall efficacy and precision of the model. 68

4.8 This confusion matrix provides a visualization of the retrieval performance at the **R@3** level of the proposed method for the **WSI-level** classification across 22 distinct primary diagnoses. It serves as a detailed display of the model's accuracy in diagnosis classification. By presenting the distribution of results across true and predicted classes, it enables a profound understanding of the overall efficacy and precision of the model. 69

4.9 This confusion matrix provides a visualization of the retrieval performance at the **R@5** level of the proposed method for the **WSI-level** classification across 22 distinct primary diagnoses. It serves as a detailed display of the model's accuracy in diagnosis classification. By presenting the distribution of results across true and predicted classes, it enables a profound understanding of the overall efficacy and precision of the model. 70

4.10 Three randomly selected patches from three distinct Cribriform Carcinoma WSIs present in the GRH dataset. The noticeable differences in structure and the significant variations among these images illuminate the challenge of accurate classification, ultimately contributing to lower accuracy rates. . 71

4.11 Visualizing image retrieval using real-world medical descriptions extracted from the WHO dataset, which wasn't included in the original training set. Under each description, the associated primary diagnosis is displayed. To the right, the top three images retrieved by the model are presented alongside their corresponding primary diagnoses. This illustration showcases the model's ability to retrieve and match images based on textual medical descriptions accurately. 72

4.12 This figure illustrates the Recall@K@V values for various combinations of V and K. In this process, the most relevant primary diagnoses are first sorted for each patch based on a given V value. Subsequently, the Recall@K is computed for each WSI using the frequency of the most common primary diagnosis identified for that WSI. The results of this process are showcased in the figure for multiple voting thresholds (V), which include 1, 2, 3, 5, 10, 15, and 22. The selection of these thresholds is significant as the number of primary diagnoses is capped at 22. Additionally, the figure also displays these results across varying levels of Recall (K), specifically 1, 3, 5, and 10. This visualization provides a comprehensive overview of how alterations in the voting threshold and recall level influence the Recall@K@V values, thereby highlighting the nuanced interplay of these parameters in the evaluation process. 75

List of Tables

3.1 Comparison with state-of-the-art methods on MS-COCO 44

3.2 Comparison with previous methods on ARCH. 46

3.3 The effect of the number of matching steps, K, in MS-COCO dataset. . . . 46

4.1 Distribution of Primary Diagnosis: This table reports the number of each primary diagnosis in the GRH pathology dataset. This dataset encompasses 22 subtypes of breast cancer, establishing it as the most extensive and comprehensive breast cancer dataset compared with literature that predominantly considers two subtypes. 58

4.2 A comparison between the proposed method and other state-of-the-art approaches, examining their performance in both patch-based and WSI-based retrieval tasks on the GRH dataset. 62

4.3 Comparative analysis of methodological performance on the GRH dataset using Recall@K@V metric with a "V" threshold of five - illustrating the superior performance and adaptability of the proposed LILE+H-DINO method across diverse recall thresholds. 73

4.4 A comparison of the proposed method and other state-of-the-art approaches on the PatchGastricADC22 dataset, evaluating their performance in patch-based and WSI-based retrieval tasks. 77

4.5 A comparison of the proposed method and other state-of-the-art approaches on the LC25000 dataset, evaluating their performance in patch-based retrieval tasks. 80

List of Abbreviations

AI Artificial Intelligence 47

CNN Convolutional Neural Network 8, 9, 17, 21

CP Computational Pathology 24, 39

CT Computed tomography 84

CV Computer Vision 3, 8, 17, 18, 28

DeiT Data-efficient image Transformer 17

EMA exponential moving average 55

GRU Gated Recurrent Unit 45

H&E hematoxylin and eosin 24, 57

HIPAA Health Insurance Portability and Accountability Act 59

LILE Look In-Depth before Looking Elsewhere xi, 26–28, 43–45, 61, 63, 64, 74–77, 79, 80, 83

LLM Larg Language Model 24, 25

LSTM Long-Short Term Memory x, 9, 10

MIL Multiple Instance Learning 21

ML Machine Learning 18, 23, 27, 48

MRI Magnetic resonance imaging 84

NLP Natural Language Processing 3, 8, 10, 17, 18, 28, 29

OOV Out-of-vocabulary 12

RNN Recurrent Neural Network 8, 9, 21, 24

SOTA state-of-the-art 8, 18, 27, 44, 47, 48, 77

SSL Self-Supervised Learning 6, 7, 18, 19, 25, 47–49, 54, 61, 63, 78–80, 82–84

ViT Vision Transformer xi, 17, 19, 29, 30, 61, 78

VLP Vision-Language-Pre-Training 23

WSI Whole Slide Image 5, 23, 24, 47–51, 57–64, 71, 73, 74, 76–78, 82, 83

Chapter 1

Overview

1.1 Motivation

The rapid expansion in the volume and variety of data, along with the recent success of deep learning, has given rise to numerous real-world applications that depend on multimodal data integration [120], [119]. These applications are often complex and require data from multiple sources to provide a comprehensive and holistic representation of data for downstream tasks. However, most existing models are limited to processing a single data modality, such as analyzing images or text exclusively. This limitation restricts the models' capacity to achieve more generalizable, robust, and versatile data representations, which is essential in today's data-driven world.

As the demand for models that fluently operate across multiple modalities escalates, the ability to effectively harness information from disparate data sources is becoming increasingly pivotal in both academic and industry settings. Such multifaceted models provide the advantage of processing diverse data types, resulting in more versatile and robust knowledge synthesis. This not only enriches the decision-making process in computer-aided systems but also catalyzes the delivery of more precise results and elevated efficiency across an array of sectors.

Multimodal learning delineates the process of architecting models capable of assimilating and correlating data from a variety of data types, or "modes", such as text, images, or audio, to name a few [20]. The overarching aim of multimodal learning revolves around creating models adept at understanding, interpreting, and linking data across diverse formats, thereby cultivating more holistic insights. The essence of this learning paradigm lies in the

fusion of these diverse modes, often yielding deeper and more intricate interpretations compared to those derived from examining each modality in isolation.

Conversely, cross-modal retrieval mainly involves using a query from one format to find items in another, like using text to retrieve relevant images. While cross-modal retrieval is a subset of multimodal learning, it faces specific challenges, like addressing the differences between these formats. Both multimodal learning and cross-modal retrieval handle diverse data, but they target different aspects of analyzing and merging such data for informed decisions.

The notion of cross-modal retrieval has rapidly gained traction in the realm of multimodal data processing. It seeks to discern the inherent linkage between data pairs originating from distinct sources, thereby unveiling the potential of blending information across various modalities. The formulation of efficient and precise cross-modal retrieval models is paramount for a multitude of applications, ranging from medical diagnosis, and content-based recommendations to multimedia search engines and beyond.

Designing and training cross-modal networks pose unique challenges, as they must simultaneously learn *inter-modal* correspondences and individual modality representations [34], [89]. Image and text are the most prevalent and commonly encountered modalities. For instance, images may be embedded within text, accompanied by captions, or referenced in textual content. In such cases, an image and its associated description often convey complementary or even redundant information. By harnessing this multimodal data, cross-modal retrieval models can offer a more comprehensive and nuanced context, ultimately enhancing the efficacy of information retrieval tasks. Furthermore, by integrating information from diverse data sources, these models can provide more generalizable and robust problem representations, facilitating better decision-making and more accurate outcomes across various downstream tasks.

Human cognition may struggle to fully understand the context by considering only the image or text. Additional information from another modality can provide more clues for identifying and analyzing the data [127]. For example, consider an image-caption pair, where the caption contains both complementary and redundant information about the image. Figure 1.1 gives an example where some parts of the caption have complementary information regarding the image; in contrast, the other parts are redundant information. This additional information aids in better understanding the image and its context. Conversely, the image can help to comprehend the context at a higher level. This example demonstrates that combining knowledge from multiple sources enhances information comprehension and retrieval performance.

Furthermore, having a model that can adeptly derive significant embeddings from

sparsely labelled data is of immense value in circumstances where acquiring labelled data is costly or presents a considerable challenge [23]. This is especially relevant in the visual domain, where the inherent complexity of the data and the obstacles associated with obtaining accurate labels make the process of extracting meaningful and informative embeddings an arduous endeavour. As a result, the advancement of cross-modal networks that can effectively learn from diverse data sources and distill valuable embeddings with minimal reliance on labelled data emerges as a crucial area of research. This field holds vast potential for a wide array of applications, encompassing medical diagnostics, Natural Language Processing (NLP), Computer Vision (CV), and beyond. The development of such models can have a transformative impact on the efficacy and efficiency of data-driven decision-making processes across numerous industries and disciplines.

The prevalence of diverse information sources and the scarcity of labelled data are especially evident in the medical domain [74]. Within patients' medical records, images, textual reports, and molecular data each serve as unique and vital sources of information, contributing to a complete understanding of a patient's condition. In this context, developing a model that can effectively extract meaningful embeddings from unlabelled data becomes of paramount importance [19]. Such a model would have the ability to retrieve and integrate information from one data source when another is available, enabling healthcare professionals to obtain a more comprehensive view of a patient's medical history, current status, and potential treatment options. This enhanced understanding would, in turn, facilitate better-informed decision-making and ultimately lead to improved patient outcomes. Furthermore, the development of these models could help address the challenge of processing and interpreting the ever-growing volume of medical data, streamlining the diagnostic process and potentially uncovering previously unrecognized patterns or relationships within the data [75]. Overall, the advancement of models capable of extracting valuable embeddings from diverse, unlabelled data sources holds significant potential for revolutionizing the medical field and improving patient care.

Cross-modal data retrieval and fusion have proven to be valuable in various domains such as healthcare, e-commerce, and education. Developing robust and efficient cross-modal models is critical to unlocking the full potential of multi-modal data, enabling enhanced decision-making and improved outcomes in numerous real-world scenarios. With the ongoing exponential growth of diverse data types, there's an urgent necessity for models that can handle and integrate this data. This underscores the need for advancements in cross-modal retrieval techniques.



Figure 1.1: Example for image-caption pair. The caption for this image reads: "*A timed exposure of the first Space Shuttle mission, STS-1, at Launch Pad A, Complex 39, turns the space vehicle and support facilities into a night-time fantasy of light. To the left of the Shuttle are the fixed and the rotating service structures*".

1.2 Objective

The common solution to understand the relationship between image and text is to map the visual semantic embeddings [79], [58] of an image and the corresponding words, phrases, sentences into a *common latent embedding space* [22], [26], [25], [128], [43], [79], [94]. In these methods, the goal is generally to find a common space in which the corresponding representations of image-text pairs are as close as possible, hence making the recognition of their relationship easier.

To gain a better understanding of the similarity between the two modalities, recent studies investigate the use of the attention mechanism. The term "*attention*" refers to a process that simulates cognitive attention to highlight the most relevant features of the input data while fading the rest. Attention-based topologies assume that the model, generally an artificial neural network, should dedicate additional computational resources to a significant portion of the data. Which part of the data is more significant than others is determined by the context and can be learned by gradient descent on training data [143]. The majority of studies in this category have employed cross-attention mechanisms, which allow the model to selectively attend to the parts of an instance that are relevant to the

context from the other modality [89], [72], [131]. Some methods attempt to refine their representation regarding the information of the other modality [34].

Nonetheless, these methods may not be able to find the optimal representation due to the semantic gap between the representation of images and text. Moreover, ambiguity in the text is a common challenge posing a learning obstacle to the model if it uses an injective embedding [131]. Aside from that, humans use a hierarchical structure to organize and store diverse semantic concepts. However, the majority of the currently available approaches group semantics together in a consistent manner [89], [131]. An approach for observing semantic concepts in a hierarchical framework that captures low-level features first and, subsequently, higher-level features could mimic human intuition more realistically. Another challenge, as mentioned in Section 1, is the difficulty and expense associated with obtaining *labelled data*. This makes extracting meaningful embeddings a considerable hurdle. To tackle this issue, some methods have explored the self-supervised learning paradigm [38]. The majority of these approaches define a *pretext task* based on unlabeled inputs to generate informative and comprehensible representations [67], [62]. However, designing these pretext tasks requires a different approach due to the distinct nature of histopathology data and its formatting as Whole Slide Image (WSI).

This dissertation addresses the mentioned problems by introducing an iterative regime that not only captures related information from the other modality but also extracts the most significant attributes by considering their context. This approach enables the model to simultaneously capture information relevant to both modalities, leading to the extraction of richer latent embeddings for each instance. Moreover, training the model iteratively allows it to gradually capture higher-level features based on those acquired in previous steps. Refining the extracted features through the intersection between modalities further enhances the model's representations.

To overcome the obstacle of labelled data scarcity, particularly for gigapixel images, this thesis proposes a self-supervised approach utilizing a novel patching scheme. This scheme aims to adapt to the intrinsic features of WSI and assist the model in generating more robust embeddings for the visual component. The entire model is trained using an end-to-end regime, which improves the accuracy and efficiency of the model [123], [88].

Extensive experiments have been carried out on public datasets MS-COCO [93], ARCH [59], and PatchGastricADC22 [142], as well as a private dataset from GRH (Grand River Hospital, Kitchener, ON) to evaluate the effectiveness of the proposed method. The ARCH, GRH, and PatchGastricAD22 datasets consist of histopathology images along with their corresponding captions/reports/primary diagnoses, which are extracted from histopathology textbooks, the PubMed repository [9], originate from clinical practice in GRH case.

1.3 Thesis Outline

The remainder of this thesis is organized as follows: Chapter 2 commences with an overview of essential background material and key concepts employed in this study. Subsequently, it delves into a comprehensive review of pertinent research papers on cross-modal retrieval and self-supervised learning, highlighting the significance of these methodologies in the current state of the field. In addition, this chapter briefly explores previous research efforts that have investigated transformers, shedding light on their relevance and impact.

Moving forward, Chapter 3 introduces the proposed approach for cross-modal data retrieval, detailing the various components and techniques employed in the model's design. This chapter also offers dataset descriptions, and a quantitative assessment of the model's performance, elucidating its effectiveness in meeting the challenges of cross-modal retrieval.

In the subsequent section, Chapter 4, the application of Self-Supervised Learning (SSL) is introduced as a means to acquire enriched and robust vision embeddings to tackle the scarcity of labelled data. When these embeddings are amalgamated with the proposed cross-modal retrieval architecture, an enhancement in retrieval performance is observed.

In addition to introducing these innovative techniques, a comprehensive description of the datasets utilized in this study is provided within this chapter. Moreover, a meticulous quantitative evaluation of the model's performance is conducted, using a variety of metrics to highlight the efficacy of the proposed approach in addressing the intrinsic challenges of cross-modal retrieval. The rigorous experimental methodology and statistical results are outlined, effectively demonstrating the superior performance of the model.

Finally, Chapter 5 presents a summary of the essential findings and insights derived from this research. In addition, it outlines potential future directions for exploration, underscoring the continued importance and relevance of cross-modal retrieval and self-supervised learning in the ever-evolving landscape of machine learning and medical imaging.

Chapter 2

Background and Related Works

This chapter attempts to provide a comprehensive overview of the main concepts used in this dissertation, as well as different types of approaches for the cross-modal retrieval tasks. To this end, this chapter has been organized into three sections. Section 2.1 begins by laying the groundwork for the proposed methodology by providing a clear and concise overview of the basic concepts of attention and Transformer-based models. This section will serve as a foundation for understanding the key ideas presented later in Chapter 3. In Section 2.2, a comprehensive overview of SSL will be provided, a powerful paradigm that has gained increasing attention in the field of deep learning. The basic concepts and principles that underlie SSL will be introduced, including its key advantages and limitations. Next, the differences between SSL and other popular learning paradigms, such as supervised and unsupervised learning, will be explored, highlighting the unique advantages of SSL, particularly in scenarios where labelled data is scarce or expensive to obtain.

The different types of SSL approach that have been proposed in the literature, including contrastive learning, generative modelling, and clustering-based methods, will be reviewed. For each approach, a detailed explanation of its underlying principles, as well as its potential applications in the context of cross-modal retrieval, will be provided.

In Section 2.3, the existing literature on cross-modality retrieval is reviewed with a specific focus on the use of attention-based approaches. The intention of this analysis is to highlight the strengths and weaknesses of these approaches and to identify potential areas for improvement.

Lastly, in Section 2.4, a brief background on digital pathology is provided, along with the specific challenges posed by cross-modality retrieval tasks in histopathology. By under-

standing this domain's unique complexities, the potential impact of the proposed methodology and its potential applications in clinical settings can be better appreciated.

2.1 Transformer Models

In recent years, the use of *Transformer* models has become increasingly popular due to their ability to outperform their predecessor models across a wide range of tasks, as substantiated by an expanding body of research [124], [96], [87], [28], [13]. These models are a type of neural network originally developed to solve sequence-to-sequence problems in machine translation [14]. However, they have since been shown to have much broader applications, as they can now be applied to any task that involves transforming one sequence into another, regardless of the modality [92], [99].

The defining characteristic of Transformer models is their **attention mechanism**, which allows them to selectively focus on different parts of the input sequence at each computation step. This mechanism has been shown to be highly effective in capturing long-range dependencies and relationships between different elements of the input sequence, which can be crucial for many tasks in both CV and NLP domains [50], [28], [54].

In the field of NLP, Transformer models have achieved and surpassed previous state-of-the-art (SOTA) performance on a wide range of tasks, from language modelling and sentiment analysis to question answering and machine translation [50], [111], [141], [45]. More recently, Transformer models have also been successfully applied to CV tasks, such as image classification and object detection, by adapting the attention mechanism to operate over 2D image grids [54], [83].

Overall, the success of Transformer models across multiple domains and tasks is a testament to their versatility and effectiveness. As researchers continue to explore new ways to leverage their power, even more exciting advances in the field of deep learning will likely be seen in the coming years.

Transformer models represent a major breakthrough in the field of deep learning, as they have shown remarkable performance on a wide range of tasks that involve processing input sequences and producing output sequences. Prior to the emergence of Transformers, Recurrent Neural Network (RNN)s and Convolutional Neural Network (CNN)s were the most popular models for sequence processing tasks, such as language modelling and speech recognition [133], [166]. However, these models suffer from certain limitations making them difficult to train and limiting their performance.

For instance, RNNs are notoriously difficult to train due to the vanishing gradient problem, which arises when gradients propagate through many time steps in the sequence, causing them to become very small and ultimately leading to poor performance [48]. To address this problem, researchers introduced LSTM networks, which use gating mechanisms to selectively retain or forget information over long time horizons [63]. LSTMs have been shown to be more effective than traditional RNNs at capturing long-term dependencies in sequences, but they still struggle when the input sequences are excessively long, as the likelihood of retaining context drops dramatically as information gets farther from the current entity being processed.

As depicted in Figure 2.1, LSTMs employ multiplication and addition to make minor changes to the input data. They use a combination of cell states and gating mechanisms to selectively retain or forget information over long time horizons. While LSTMs have been successful in many sequence processing tasks, their limitations in dealing with excessively long sequences are a big problem. This problem arises as the likelihood of retaining context drops dramatically as information gets farther from the current entity being processed [2]. In addition to the vanishing gradient problem, RNNs and LSTMs suffer from another major limitation, which is their sequential processing nature. This makes it difficult to parallelize the computations, which can significantly slow down training and inference time [56].

In contrast to RNNs and LSTMs, which suffer from the vanishing gradient problem and have difficulty processing excessively long input sequences, CNNs have been successful in exploiting local dependencies. They can be parallelized, allowing for efficient processing of large amounts of data. This is because CNNs operate on small, localized regions of the input and do not require knowledge of the entire sequence to make predictions. Consequently, CNNs have been used in many applications, such as image classification, object detection, and natural language processing.

Despite their strengths, CNNs do not necessarily solve the problem of capturing *long-range dependencies* in sequences. This is because they are designed to operate on fixed-size input tensors and are typically applied in a feed-forward manner, which does not allow for feedback loops between different parts of the network. As a result, CNNs can struggle to capture more complex patterns in sequences, particularly those that involve longer-term dependencies between different elements [54].

Recent advances in deep learning, particularly the development of Transformers, have offered a promising solution to these limitations. Transformer models use self-attention mechanisms to selectively focus on different parts of the input sequence, allowing them to capture long-range dependencies more effectively and process input sequences in parallel.

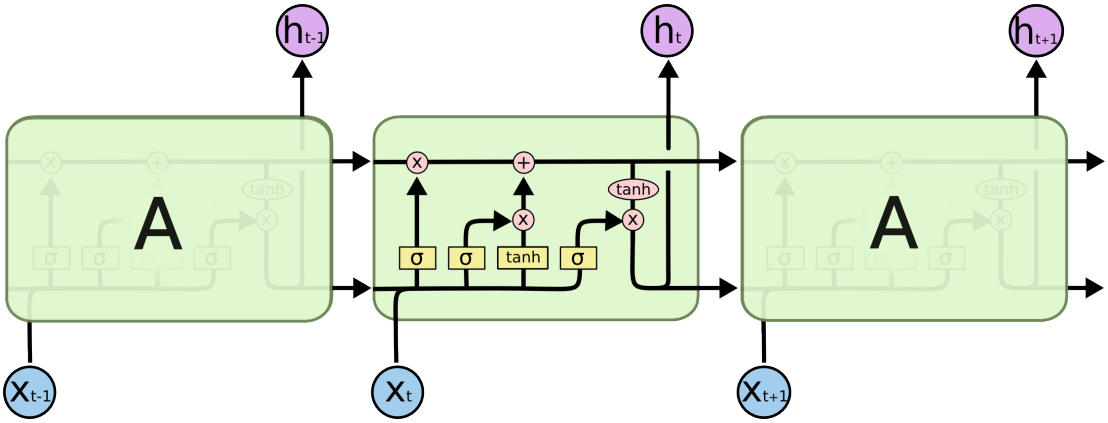


Figure 2.1: An LSTM's repeating module is organized into four interacting layers. Each line contains a full vector from one node's output to the inputs of others. The pink circles indicate point-wise operations such as vector addition, and the yellow boxes represent learned neural network layers. Lines that merge indicate concatenation while forking denotes their content being copied and the copies being sent to distinct destinations. Figure adopted from [11]

.

To solve the aforementioned problems, the Transformer concept was introduced in the paper "Attention is all you need" [143], which proposed a novel way to perform sequence translation in NLP applications. As depicted in Figure 2.2, the Transformer architecture is composed of two main components: an encoder on the left and a decoder on the right. Both the encoder and decoder consist of blocks that can be stacked on top of each other multiple times. Each block in the Transformer architecture is based on two key components: a multi-head self-attention module and a feed-forward layer.

The input to the Transformer first flows through an input embedding layer, which converts the given inputs into embedding vectors. This process is described in detail in Section 2.1.1. These embedding vectors are then added to a positional embedding, which provides the Transformer with information about the order of the input and how the input sequences are fed to the network [143]. This step is explained in Section 2.1.2. After the input sequences are transformed into embedding vectors and positional embeddings, they are fed into a stack of encoder blocks. Each encoder block consists of a multi-head self-attention module, which allows the model to selectively attend to different parts of the input sequence, followed by a feed-forward layer that applies a non-linear transformation to the outputs of the attention module. The encoder block is explained in Section 2.1.3.

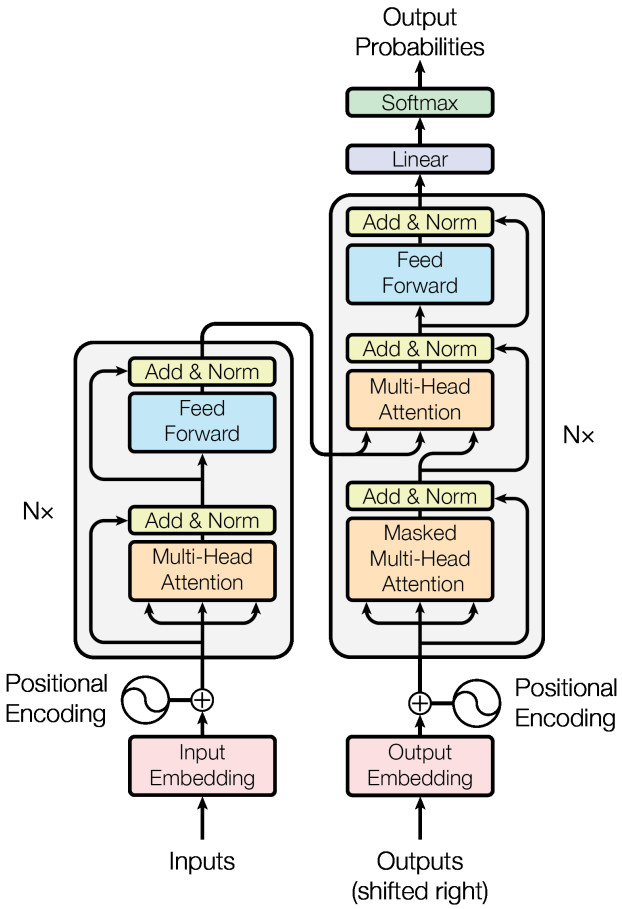


Figure 2.2: The Transformer architecture. The main components are the encoder and decoder parts. The embedding layer and positional embedding are present in both the encoder and the decoder. Aside from that, the Encoder stack is composed of several encoders. Each encoder has a multi-head attention module as well as a feed-forward layer. The final encoder is fed into the decoder stack, which comprises a number of decoders. Each decoder has two multi-head attention layers as well as a feed-forward layer. Image adapted from [143]

2.1.1 Input Embedding

The input embedding layer plays an important role in the Transformer architecture by transforming each entity in the input sequence into a dense, continuous vector representation. The input embedding layer in natural language processing can be viewed as a *lookup*

table that retrieves a learned vector representation for each token[1] in the input sequence.

The embedding layer associates each token with a unique, d-dimensional vector representation, enabling the model to capture the semantic and syntactic relationships between words in the input sequence. Figure 2.3 demonstrates the transformation of a sequence of tokens into their respective embedding vectors. During training, embedding vectors are learned and can be optimized to capture various input sequence characteristics, such as word meanings, sentence structure, and context.

Out-of-vocabulary (OOV) words, which are words that do not appear in the model's predefined vocabulary, can be effectively handled by the model when input embeddings are used. By representing each word as a continuous vector, the embedding layer can capture the similarity between OOV words and other words in the vocabulary, enabling the model to generalize to new words and phrases that may not have been encountered during training.

As a result, the input embedding layer is an integral part of the Transformer architecture that enables the model to effectively process input sequences by transforming each entity into a dense, continuous vector representation. By mapping each token to a unique embedding vector, the model is able to capture the semantic and syntactic relationships between words in the input sequence, as well as handle out-of-vocabulary words effectively.

2.1.2 Positional Embedding

When processing a sequence, the order of the elements is crucial for capturing dependencies and relationships between different parts of the sequence. However, when the sequence is transformed into a set of discrete tokens, the model must rely solely on the learned embeddings to capture this information. To address this issue, the Transformer architecture uses positional encoding to provide the model with additional information about the position of each token in the sequence. There are two main types of positional encoding used in the Transformer: *absolute* positional encoding and *relative* positional encoding.

The authors of the original Transformer paper, "Attention is all you need" [143], proposed the use of a sinusoidal function for absolute positional encoding in the Transformer architecture. This approach assigns a unique wavelength to each position in the input sequence, encouraging the model to focus on a specific range of positions. This function encourages the model to focus on a specific wavelength, λ, regarding the position of each

[1] *Tokenization* is the process of breaking down a piece of text into small units called tokens. A token may be a word, part of a word or just characters like punctuation.

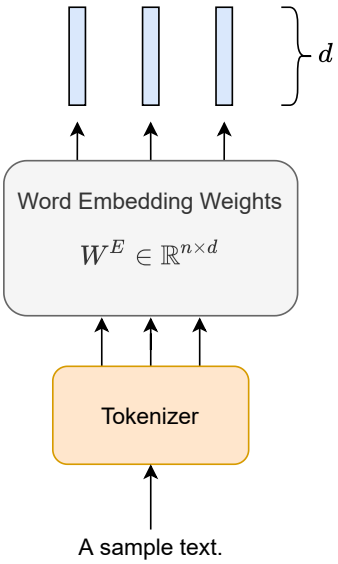


Figure 2.3: The Embedding layer. The word embedding layer will map each token to a vector spanned in d dimensions (the embedding dimension). n is the number of entries in the dictionary, and d is the vector dimension.

token. Given a signal $y(x) = sin(kx)$ the wavelength is $\frac{2\pi}{\lambda}$. In this case, λ is determined by the position in the input sequence. The running variable i is employed to differentiate between odd and even positions. The positional embedding for each token can be derived from Eq. 2.1, where d is the size of the embedding vector.

$$\begin{aligned} \mathrm{PE}_{(pos,2i)} &= sin\left(\frac{pos}{10000^{2i/d}}\right) \\ \mathrm{PE}_{(pos,2i+1)} &= cos\left(\frac{pos}{10000^{2i/d}}\right) \end{aligned} \tag{2.1}$$

For example, for token t at position $pos \in [0, S-1]$ in the input sequence $t = (t_0, \ldots, t_{S-1})$ with 4-dimension embedding vector, the positional embedding would be

$$\mathrm{PE}_{(pos)} = [sin(\frac{pos}{10000^0}), cos(\frac{pos}{10000^0}), sin(\frac{pos}{10000^{\frac{2}{4}}}), cos(\frac{pos}{10000^{\frac{2}{4}}})]. \tag{2.2}$$

The value of 10000 in the original paper seems to have been chosen somewhat empirically to strike a balance between capturing fine-grained position information and avoiding overfitting to specific positions in the sequence.

In contrast, relative positional encoding uses the relationship between different tokens in the sequence to encode positional information [125]. This method is based on the fact that the relationships between different tokens in the sequence can be shown as a matrix, which can be used to derive a set of relative positional embeddings. These embeddings capture the pairwise relationships between different tokens in the sequence and can be added to the learned embeddings to provide the model with additional positional information. The relative position between two tokens is represented using a learned embedding vector $\mathbf{r}_{i-j}$, where i and j are the indices of the two tokens being compared. This relative position embedding is added to the dot product between the query and key vectors in the self-attention mechanism as follows

$$\mathbf{A}_{i,j} = \text{softmax}\left(\frac{\mathbf{q}_i \times (\mathbf{k}j + \mathbf{r}i - j)}{\sqrt{d}}\right) \tag{2.3}$$

where $\mathbf{q}_i$ and $\mathbf{k}_j$ are the query and key vectors corresponding to the ith and jth tokens, and d is the dimensionality of the query and key vectors. The relative position embedding $\mathbf{r}_{i-j}$ is learned during training and allows the self-attention mechanism to capture the relative position of different tokens within the input sequence. This approach has been shown to be effective for tasks that require modelling long-range dependencies and capturing complex relationships between different parts of the input sequence.

The use of both absolute and relative positional encoding in the Transformer architecture enables the model to effectively capture the order of the input sequence and, thus, more accurately capture the underlying relationships between the sequence's various elements. By incorporating both types of positional encoding into the input sequence, the model is able to capture the absolute position of each token as well as the pairwise relationships between tokens and utilize this information more effectively to enhance the quality of its predictions.

2.1.3 Encoder

The Transformer encoder is a critical component that plays a key role in transforming an input sequence into a continuous, abstract representation that captures the relevant information for the entire sequence. As shown in Figure 2.2, each encoder layer is composed of a multi-head self-attention module and a feed-forward layer, with layer-normalization and residual connections employed around each sub-layer to facilitate gradient flow and network normalization [143].

2.1.4 Attention

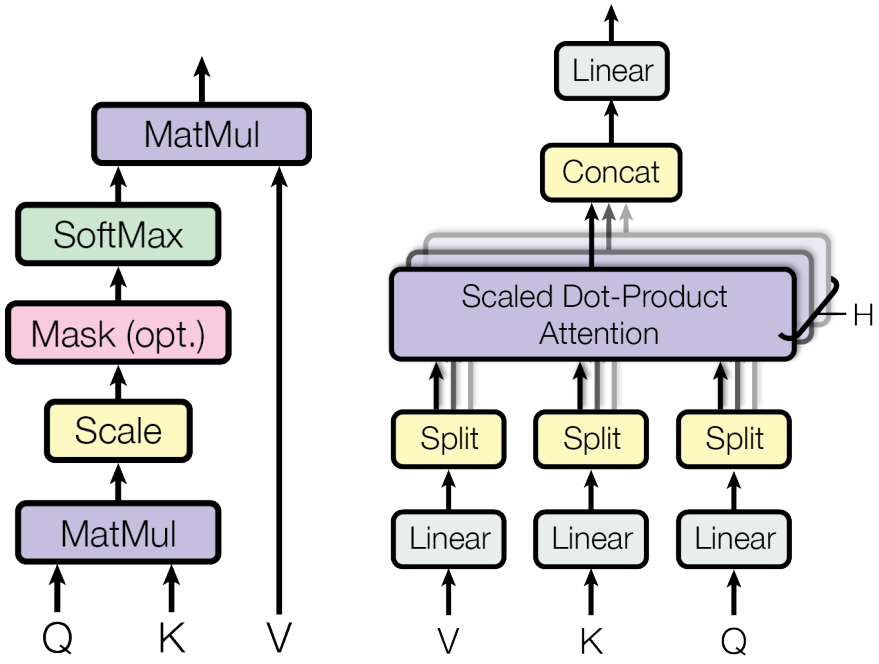
The Transformer has gained considerable traction in recent literature due to its efficient utilization of the *attention* mechanism [135]. Designed to aid neural machine translation in memorizing lengthy source texts [135], [118], the attention mechanism allows the model to create shortcuts between the context vector and the complete source input. This is achieved by applying attention instead of constructing a single context vector from the encoder's prior hidden state [151]. Each output is a weighted sum of values, with the weight allocated to each value based on the similarity of the query with the corresponding key [159].

Additive attention [17] and dot-product attention are the most prevalent attention functions. In the original Transformer paper [143], the authors recommend employing scaled dot-product attention, as it aligns more closely with their algorithm. Scaled dot-product attention proves to be faster and more space-efficient compared to alternative approaches such as additive attention [17]. Figure 2.4 visually represents the attention module's components. The attention function in practice can be illustrated by Equation 2.4. In this equation, Q represents a set of queries and K corresponds to a set of keys, both possessing dimensions of d_k. Simultaneously, V represents the values, which have a dimension of d_v.

$$\textbf{Attention}(Q, K, V) = \textbf{softmax}\left(\frac{Qk^T}{\sqrt{d_k}}\right)V. \tag{2.4}$$

This concept intuitively, suggests that multiple heads allow the model to focus on distinct segments of a sequence separately [47]. Instead of applying a single attention function to d_{model}-dimensional data, the multi-head attention model performs linear projections on queries, keys, and values h times in parallel. These projections are individually learned, thus enabling the derivation of the dimensions d_k, d_k, and d_v respectively. As a consequence, the dimensions for d_k and d_v can be calculated by $d_k = d_v = \frac{d_{model}}{h}$, revealing the intrinsic relationship between the dimensions and the number of heads in the model [143]. The attention function is subsequently applied in parallel to each of the projected versions of the queries, keys, and values, generating d_v-dimensional output values. These steps, demonstrated in Figure 2.4, are added together after they have been computed.

Another rationale for utilizing multi-head attention is that it allows the model to attend to input from distinct representation sub-spaces at different points while averaging in single-head attention prevents this [143].



(a) (b)

Figure 2.4: (a): Scaled dot-product attention module. (b): Multi-head attention. The attention module performs its computations in parallel many times. Each of these is referred to as an *attention Head*. The attention module divides its Query, Key, and Value arguments N times and feeds each split via a separate Head. The results of all of these identical attention computations are then added to generate a final attention score. Figures adapted from [143]

$$\textbf{Multi-Head}(Q, K, V) = \textbf{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O \tag{2.5}$$

where,

$$\text{head}_i = \textbf{Attention}(QW_i^Q, KW_i^K, VW_i^v) \tag{2.6}$$

Here, Ws are learnable parameter matrices with dimension of $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^v \in \mathbb{R}^{d_{model} \times d_v}$ and $W_i^O \in \mathbb{R}^{hd_v \times d_{model}}$

2.1.5 Transformer Models in Computer Vision

Transformer architecture breakthroughs in the NLP tasks have inspired considerable interest in the CV field in adapting these models for the vision and multi-modal learning problems. Nevertheless, visual data has a consistent structure (e.g., spatial and temporal coherence), necessitating unique network designs and training approaches. Various adaptations and extensions of Transformer models have been proposed to tackle different computer vision challenges including image identification [54], [139], object detection and segmentation [30], [170], [156], image super-resolution [155], image generation [33], text-image synthesis [120], and visual question answering [134], [132], among other applications [147], [86], [99]. Some of the most prominent vision Transformer models that have made a significant impact on the field are reviewed in this section.

ViT [54] was the first implementation of a Transformer for computer vision tasks. It applied the original Transformer blocks with some minimal modification to the input data. The authors suggest that instead of a sequence of token embeddings fed to the network in a standard Transformer, the input image is reshaped to a sequence of flattened patches. ViT requires a larger dataset to be trained compared to CNNs since CNNs encode prior knowledge (e.g., translation equivariance) about the images, which allows them to reduce the amount of data that they require. [82]. The Data-efficient image Transformer (DeiT) paper [139] was the first to show that Transformers can be learned on mid-sized datasets as well (1.2 million ImageNet [49] examples against 300 million images used in ViT). The primary contribution of DeiT is a novel native distillation strategy for Transformers to leverage a CNN as a teacher model to train the Transformer model, in addition to applying augmentation and regularization processes prevalent in CNNs. The CNN outputs in the DeiT architecture assist the Transformer in determining valuable representations for given images. The Swin Transformer proposes a hierarchical Transformer model with shifted windows [97]. It divides an image into non-overlapping patches and processes them hierarchically, allowing for better scalability and performance on a wide range of vision tasks, including image classification, object detection, and semantic segmentation. Another utilization of Transformer architecture for vision is ConViT (Convolutional Vision Transformers) [55]. The author introduces a convolutional-like attention mechanism that combines the benefits of both CNNs and Transformers. This approach incorporates local information processing with the global receptive field of Transformers, resulting in improved performance on computer vision tasks. These examples represent a few of the many innovative adaptations and extensions of Transformer models for computer vision tasks. The ongoing development of new methods highlights the potential of Transformer-based models to revolutionize the field of computer vision further. There are several other Transformer

based models that are implemented for CV applications [82], [163], [21]. Nevertheless, the main idea is to apply Transformer blocks in their architecture.

2.2 Self-Supervised Learning (SSL)

In recent years, SSL has emerged as an effective method in the field of Machine Learning (ML) and has drawn immense interest. SSL operates on unlabeled data, which is abundant and readily available for many applications, in contrast to supervised learning, which requires a significant amount of labelled data. SSL is designed to learn effective representations of data without the need for explicit supervision by learning useful features from the data itself. Models trained with SSL can be fine-tuned for a wide range of downstream tasks, often with less labelled data than would be required for purely supervised approaches.

The primary goal of SSL is to enable machines to learn in a more human-like manner by leveraging the vast amounts of unannotated data that humans ingest and learn from with ease. This learning paradigm has been implemented in numerous fields, including NLP, CV.

SSL has proven to be particularly effective in CV, where it has achieved SOTA performance in many tasks, including image classification, object detection, and semantic segmentation [112], [154], [16]. Self-supervised methods in CV usually utilize pretext tasks, which are designed to generate surrogate supervision signals from the raw data itself in order to learn feature representations. Image inpainting, image colorization, and image rotation prediction are prominent pretext tasks [52], [60], [39]. There are some popular methods in self-supervised learning that in this section are going to elaborate them. SimCLR [37], BYOL [64], and DINO [32] are three popular self-supervised learning methods in the field of CV. Each of these methods uses a different approach to learn effective representations from unlabeled data. SimCLR (Simple Framework for Contrastive Learning of Representations) is a method for learning representations by maximizing the agreement between various augmented views of the same data point. SimCLR employs a "projection head" to transfer the representations of the augmented views to a common feature space in which agreement between the views can be maximized. The key idea of SimCLR lies in its utilization of contrastive learning to acquire meaningful representations. Additionally, the incorporation of a large batch size and a temperature scaling parameter has been shown to enhance the stability and performance of the method. BYOL (Bootstrap Your Own Latent) is an approach that employs a "predictor" network to predict the representations of a set of augmented views of the same data point. The predictor network is trained to

generate the representations of one view, while the "target" network is used to generate the representations of another view. Optimizing the predictor network to match the representations produced by the target network encourages the target network to learn more informative representations. The primary innovation of BYOL lies in its adoption of SSL without negative pairs during training, combined with the incorporation of an exponential moving average (EMA) mechanism to update the target network. These novel approaches contribute to enhanced stability and faster convergence during the training process. DINO (DIstillation of knowledge with NO labels) is a SSL method that employs self-distillation and ViT to acquire high-quality visual representations without labels. The main innovation of the DINO approach lies in the application of "self-distillation", where both teacher and student networks share an identical architecture. The method optimizes the student network to mimic the output of the teacher's network, which is achieved by maintaining an exponential moving average of the student network's weights. This unique self-distillation process that introduced in [68] enhances the student's ability to learn meaningful representations from the data, leading to improved performance and generalization. Moreover, this approach effectively mitigates mode collapse, a common issue in deep learning, by encouraging the student network to explore a wider range of representations during training. Refer to Figure 2.5 for a graphical representation of the DINO model.

2.3 Cross-Modality Retrieval

This study aims to investigate cross-modal image-text retrieval, which involves finding the underlying relationship between images and textual descriptions. Cross-modal retrieval is a well-studied problem in multimedia research and has several applications, including image captioning, text-based image retrieval, and visual question answering. In recent years, deep learning-based methods have shown significant improvements in cross-modal retrieval tasks. This section will explore the most relevant studies on image-text retrieval.

A rich line of works on cross-modality retrieval was looking for a similar pattern in the feature representation of both modalities [79], [89], [73], [100], [91]. These studies can generally be categorized into global and local feature-matching approaches. Global matching methods attempt to map whole images and complete words to a shared feature space. On the other hand, local feature-matching approaches make use of local features to attempt to match image regions with words.

A substantial body of research on cross-modality retrieval focuses on finding similar patterns in the feature representations of both modalities [79], [89], [73], [100], [91]. These studies can generally be categorized into global and local feature-matching approaches.

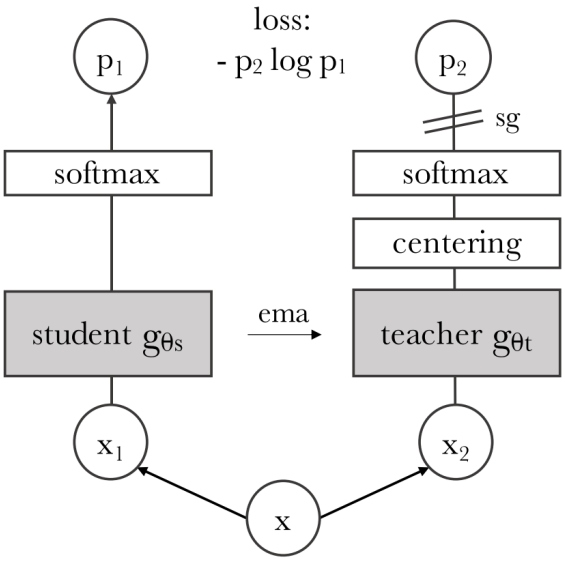


Figure 2.5: The DINO model utilizes different distinct random transformations of an input image (for simplicity, only two of them are shown in this Figure), which are respectively passed to the student and teacher networks. While both networks share the same architecture, their parameters are distinct. The output of the teacher network is centred using a batch-wide mean. Each network generates a K-dimensional feature that is normalized using a temperature softmax over the dimension of the feature. Then, a cross-entropy loss is used to determine their degree of similarity. Stop-gradient (sg) operators are applied to the teacher in order to propagate gradients only through the student. The parameters of the teacher are updated with an exponential moving average (ema) of the parameters of the student model. Figure adopted from [32].

Global matching methods attempt to map entire images and complete words to a shared feature space, while local feature matching approaches utilize local features to match image regions with words.

Global feature matching methods were among the first attempts to tackle the image-text retrieval task. Kiros *et al.* [85] proposed to utilize a CNN and a RNN to encode given images and text data. They introduced a pairwise ranking loss objective for the loss function, aiming to rank images and their corresponding descriptions. Faghri *et al.* [57] modified the loss function by incorporating hard negatives in the triplet loss function and applied it to an architecture similar to [85]. They demonstrated that adding a hard negative loss function could lead to significant improvements. In addition to these efforts, several other approaches have been proposed. The m-CNNs paper proposes multimodal convolutional neural networks (m-CNNs) for image-sentence matching. The m-CNNs consist of one image CNN and one matching CNN [104]. The authors claim that their approach can effectively capture the information necessary for image and sentence matching.

As the quality of generative models improved [169], [81], some studies attempted to exploit the capability of generative models to enhance the performance of proposed methods. Mandela *et al.* [113], Peng *et al.* [114], and Gu *et al.* [65] suggested incorporating a generative objective into the loss function, which helped bring related instances closer together.

Local feature matching methods have shown to be effective in capturing the correspondence between image and text feature representations at the level of image regions and words. These methods have gained more popularity due to their ability to match image regions and words separately, resulting in better performance compared to global feature matching methods [131]. Karpathy *et al.* proposed a method to extract features for image regions at the object level using Faster R-CNN [121] and text words and then align them into a common space [79]. As the application of the attention mechanism in a variety of tasks has become more common, several approaches in image-text retrieval seek to leverage it in their methods [34], [89], [76], [131], [136], [89]. Song *et al.* [131] proposed a method to combine global features with locally-guided features via a multi-head self-attention module. They implemented their approach by applying two identical networks for vision and text and optimized them simultaneously in a Multiple Instance Learning (MIL) framework. Thomas *et al.* [136] in a similar network proposed a novel within-modality loss function that drives feature representation toward more coherence. Their approach assures that not only paired images and text are close, but also image-image and text-text relationships are preserved. The majority of these methods are based on the assumption that vision and text are distinct and equally important modalities. However, grounding the representation

of each modality derived from the other at a finer level of detail is crucial for bridging the gap between vision and text. Lee *et al.* [89] was among the first studies that investigated the use of cross-attention to explore full latent matching using image regions and words in the text as context.

Furthermore, Chen *et al.* in IMRAM paper [34] presented an iterative alignment method that progressively captures fine-grained correspondence between image and text using a cross-attention module. They claimed that a memory unit could assist the model in refining the feature representation from early steps to later ones, which helped them to achieve impressive results. With the development of Transformer models, advanced approaches have been proposed for cross-modality retrieval tasks using Transformers [99], [134]. In [99], the authors drew inspiration from BERT's [50] success in language modelling and proposed a novel network architecture. This architecture comprises two single-modal Transformer networks applied to input images and text data, respectively. These single-modal Transformers are then followed by a cross-modal Transformer, facilitating information exchange between modalities by enabling the network to attend to the other modality. This design enables the network to leverage both image and text information effectively, resulting in a powerful cross-modal representation learning framework.

There is another line of research that has tried to encapsulate image-text retrieval into graph Convolution networks [145], [51]. Diao *et al.* [51] employed a similarity graph reasoning network to infer relation-aware similarities with both the local and global alignments. Then they applied attention filtration to integrate these alignments effectively by selectively attending to the significant and representative alignments.

Recent studies have demonstrated that training on large-scale datasets can significantly improve the performance of the models in image-text retrieval tasks [91], [116], [77]. Jia *et al.* in ALIGN paper [77] leveraged a dataset comprising over one billion paired images and alt-text data, asserting that a straightforward dual-encoder network could effectively align image and text representations using a contrastive loss function. Similar to the ALIGN method, Xiujun *et al.* in Oscar paper [91] gathered a dataset of over 6.5 million paired images and texts to train their network. Authors in Oscar's paper suggested utilizing the output of an object detector network as anchor points to align image regions with word embedding vectors. Another example of boosting the performance using a large amount of data can be CLIP [116] paper. Authors in the CLIP paper also demonstrated that the task of image-text retrieval could be outperformed using a massive dataset, 400 million in this case, and proper networks, GPT-3 [28] for the text encoder, and ResNet-50 [69] for the image encoder. However, a bottleneck of these kinds of large-dataset-dependent approaches is with those domains which are unlikely to be included in their collected dataset, such as the histopathology domain. Moreover, training on that amount of data requires huge com-

putational resources. Some other papers focused on the importance of the relation between image and text pairs and proposed a novel approach to prioritize loosely aligned samples in cross-modal retrieval tasks [137]. Unlike prior methods, authors claim that their approach estimates the semantic similarity in separate channels and uses image-text pair weights to focus on diverse or discrepant neighbourhoods that help them demonstrate significant performance gains. Recently, researchers have explored the potential of Vision-Language-Pre-Training (VLP) for aligning images and text. These approaches showed promising capabilities for a wide range of cross-modal learning tasks [162], [144], [161]. In the PTP paper [144], the authors applied a novel Position-guided Text Prompt (PTP) paradigm to enhance the visual grounding ability of cross-modal models trained with VLP. The proposed approach divides the image into blocks and reformulates the visual grounding task into a fill-in-the-blank problem given a PTP, improving the visual grounding capability of VLP models. Another paper in this direction is X-VLM [162]. Authors introduced a new method called X-VLM for multi-grained vision language pre-training that can learn relations among multiple objects by locating visual concepts in the image given the associated texts and aligning the texts with the visual concepts in multi-granularity.

2.4 Cross-Modality Retrieval in Histopathology

The digitization of tissue glass slides has recently become more common in pathology practice [46]. The expedited adoption of digital pathology over the last decade has resulted in the introduction of ML methods for analyzing WSIs [105]. The applications of ML in digital pathology include reducing pathologists' workloads and optimizing disease treatment methods [105]. The computer analysis of WSIs presents a number of difficulties due to the size and complexity of images. These obstacles need research into more efficient methods of WSI analysis. One approach that can be applied to facilitate the process of WSI analysis can be image-text retrieval. Nevertheless, the lack of large datasets that include descriptions for WSIs is evident. Moreover, most of the available datasets in the pathology domain consist of descriptions for the entire WSI and not for individual patches [3]. The size of WSIs in histopathology can easily exceed 10,000 pixels which can be challenging to process. As a result, patch-based methods can be a good alternative. However, a well-structured dataset containing WSI patches and corresponding descriptions is necessary. These issues make cross-modality retrieval in histopathology became more challenging.

The *cross-modality retrieval* tasks are new to this field. There are not many studies that explore retrieving a description correspondence to an image or vice versa. As a result, a few works were trying to retrieve or utilize image-text modalities in their suggested meth-

ods. Zhang *et al.* [167] was among the first authors fwho proposed a model and dataset for image-text in pathology. The authors proposed a new framework, namely MDNet, to establish a direct multimodal mapping between images and diagnostic reports. To be able to perform that, they implemented a novel residual network and attention mechanism in the proposed method. Moreover, they developed a private patch-based dataset based on pathology bladder cancer images. The created dataset includes 1,000 hematoxylin and eosin (H&E) stained patches that are extracted from WSIs at 20x magnification level. For descriptions, they asked pathologists to write five short sentences for each patch. Thus, they created a dataset that consists of 5 ground-truth per image in total. The proposed method is capable of generating short diagnostic reports and retrieving images based on symptom descriptions. Yet the developed dataset was too small, private and limited to only H&E images. Recently, Gamper *et al.* [59] developed a new publicly available dataset, ARCH, that contains more than 7,000 patch-based images with the corresponding captions. The authors of the paper used PubMed medical articles [9] and academic pathology textbooks to collect the ARCH dataset. They selected all PubMed journals that contain keywords like "pathology", "histopathology" or "histology". Then they extracted figures and corresponding descriptions from selected journals. Moreover, they extracted figures and captions from 10 pathology textbooks to add more data to their dataset. As a result, they managed to collect 11,816 bags and 15,164 images in total[2]. Moreover, they reported that a pre-trained model on image text learns better representations for most Computational Pathology (CP) tasks and performed some experiments in different tasks to prove their observations. Some other studies, including [142], tried to tackle the caption generation tasks in the field of pathology. These researchers established a dataset that comprises matched sets of patches and corresponding reports. They fed the condensed feature maps obtained from these patches incrementally into a RNN model, demonstrating an innovative approach to this complex task. In light of recent advancements in Larg Language Model (LLM)s and visual-language foundation models, researchers have sought to apply these methodologies to the medical domain.

One such approach is CONCH [101], a study in which the authors amassed a collection of over 1.1 million histopathology image-caption pairs from publicly available articles on PubMed. Utilizing a combination of Transformers within their network for vision, text, and multimodal decoding, they demonstrated some results across diverse downstream tasks. However one has to be skeptical whenever medical data are scraped from the Internet. The concept of "Garbage In, Garage Out" in computer science teaches us that such models may be solely a very limited toy project. In a similar vein, the authors of the BioMedCLIP

[2]There is a disparity between the number of samples within the paper and the dataset available for download. Based on the available dataset, they only have around 7,000 pairs of images and captions

paper [165] introduced an approach based on the same concept. Their work involved the collection of an astounding 15 million image-text pairs extracted from over 4.4 million publicly accessible full-text articles related to the medical field. Adapting the CLIP method [116] for the biomedical domain, they exhibited its performance in an array of downstream tasks, including cross-modal retrieval tasks. These recent studies showcase some potential of leveraging LLMs and visual-language foundation models within the realm of medical research. However, skepticism prevails in the face of scarping data from the Internet.

An innovative methodology that belongs to the category of local feature-matching techniques is introduced in this research. The quality of image embeddings is amplified by the proposed method, utilizing SSL, a strategy found to be particularly beneficial in scenarios dealing with smaller datasets, which are typical in fields such as pathology and medical imaging.

While the proposed approach is conceptually akin to previous techniques, including those presented by [34], [79], it distinguishes itself by the utilization of SSL and a composite of self and cross-attention modules and its novel architecture including loss function. The encapsulation of the most meaningful feature representations before they are fed into the similarity measurement module is the primary objective of this unique design.

By leveraging SSL for cross-modal retrieval tasks, the proposed methodology highlights the possibility of marked enhancements in the precision of these tasks. Achieving this lays the groundwork for a fresh and potentially influential method to forward the domain of machine learning in areas like pathology where data resources might be scarce.

Chapter 3

Cross-Modal Retrieval

3.1 Introduction

As stated in Chapter 1, the primary objective of this dissertation is to develop a novel methodology for image-text cross-modal retrieval that efficiently captures the inherent relationships between images and their corresponding text. This methodology utilizes an iterative regime to refine the latent embeddings, leveraging the most significant attributes of both modalities in their respective contexts. This study's aim is to enhance the attention mechanisms of the model in such a way that the important components of the data are not only identified but also the representation gap between image and text modalities is bridged, ultimately resulting in more accurate and efficient retrieval results. The model's framework is demonstrated in Figure 3.1.

The proposed model includes different components. LILE, a dual attention network using Transformers, takes images and texts as inputs and extracts feature representations for each of those using Transformers. The details of this step are described in Section 3.2. The proposed self-attention module [143], which is applied for extracting the most significant parts of each modality with respect to itself, is elaborated in Section 3.3. Then, the cross-attention module and memory block, which help the model to refine the representation of each instance with respect to the outputs of attention modules, are described in Section 3.7. Additionally, the iterative matching scheme that makes use of the memory block to assist the model in refining the extracted feature representation is illustrated in Section 3.5. Finally, section 3.6 explains the proposed loss function in detail.

The proposed methodology encompasses several key components that work collectively towards accomplishing the described objectives. The forthcoming sections will delve into

the intricate aspects of these components, providing a comprehensive understanding of their design and functionality.

Firstly, the LILE architecture is introduced. The LILE architecture is a dual-attention network that leverages Transformer architectures. This architecture is designed to handle both image and text inputs, extracting feature representations using Transformers. A detailed explanation of this step can be found in Section 3.2. Following the discussion on LILE, the self-attention module in Section 3.3 is elaborated. This module, inspired by Vaswani *et al.* [143], plays a crucial role in identifying and extracting the most significant parts of each modality in relation to itself. A thorough understanding of this self-attention mechanism is acquired, providing insights into the methodologies employed to tackle the challenge of aligning representations across various modalities. Section 3.7 offers a comprehensive overview of the cross-attention module and the memory block. These components are designed to aid the model in refining each instance's representation with reference to the outputs of the self-attention and cross-attention modules. The integration of these features in the methodology demonstrates a unique approach to bridging the representation gap between images and text. The iterative matching scheme that has been applied in the model is then discussed in Section 3.5. This scheme employs the memory block to assist the model in progressively refining the extracted feature representation. This iterative process is crucial to the methodology as it enables the model to capture higher-level features based on those learned in previous steps. Lastly, Section 3.6 provides an in-depth explanation of the proposed loss function. This loss function is meticulously designed to align with the objectives of this study, providing a coherent means of evaluating and driving the learning of the model. The distinct features of the proposed method compared to the IMRAM approach [34] include the incorporation of self-attention modules to highlight key features within each modality, and an additional loss term to enhance model robustness in representation. Furthermore, the memory block was refined and integrated an attention module, further ensuring network consistency — another innovation over the IMRAM paper [34]. By combining these elements, the suggested method seeks to set a new standard (SOTA) for image-text cross-modal retrieval tasks.

3.2 Feature Representation

Data representation, also known as representation learning, plays a crucial role in ML algorithms as it enables models to automatically discover meaningful representations from the given data [24]. Effective data representation is essential for capturing the underlying structure and patterns in the data, leading to improved performance and generalization.

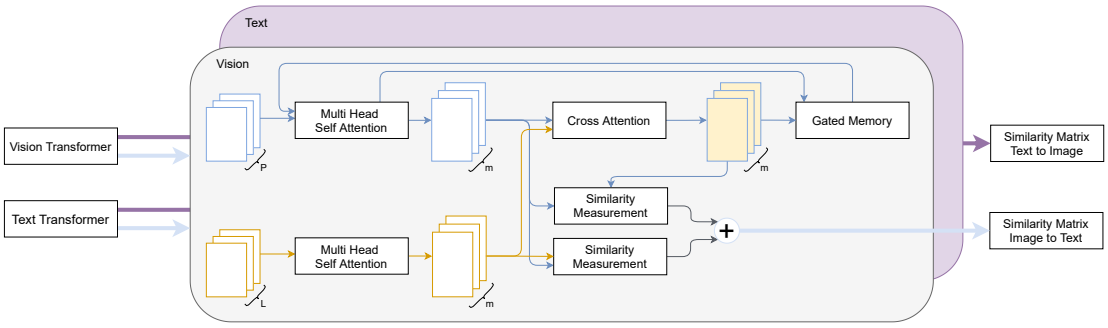


Figure 3.1: The architecture of LILE for image-text data is shown. It comprises feature extractors dedicated to each modality, with a vision transformer for images and a text transformer for texts. These feature extractors are followed by two networks with identical architecture, facilitating the integration of information from both modalities.

In recent years, deep models have demonstrated their capability to extract intricate and high-dimensional features from the data, surpassing the performance of traditional feature extractors [109], [41]. One notable class of deep models is the Transformer-based models, which have shown remarkable success in various NLP downstream tasks [50], [96], [87], [117]. The exceptional performance of Transformers in NLP has sparked interest in adapting these models for CV tasks as well [54], [139], [97], [108], [53], [153], [140].

Recent studies have explored the use of different types of Transformer models for vision tasks and have shown their effectiveness in capturing and representing visual information [95], [35]. The versatile and modular architecture of Transformer models enables them to process various modalities, including images, videos, text, and audio, by leveraging similar processing blocks. This flexibility allows for an efficient adaptation of Transformer models to different domains and tasks. Additionally, Transformers can scale effectively to handle large capacity networks for complex tasks and perform well with massive datasets.

This study adopts the Transformer architecture for extracting features from both image and text data. Transformers are particularly well-suited for processing text due to their inherent design, which enables them to capture long-range dependencies and contextual information. Additionally, the modular nature of Transformers and their ability to perform parallel processing makes them highly effective in handling the complex and diverse visual information present in images. By utilizing Transformers for text and image feature extraction, this study aims to leverage their powerful representation learning capabilities and exploit their capacity to capture intricate visual and textual relationships.

3.2.1 Image Representation

As noted earlier in this section, Transformers have been extensively used for feature extraction in tasks such as image representation. In this study, the input images are encoded using a pre-trained ViT [54]. The images are divided into a sequence of fixed-size, non-overlapping patches, and each patch is fed into the ViT to obtain a feature representation. The resulting tensor, representing the last hidden state of the ViT model for the given input, is used as the feature representation for the query images. Importantly, the length of the sequence remains constant across all layers, and the second dimension of the last hidden state corresponds to the number of patches in the input images.

Another approach for extracting visual information from images is the utilization of object detection models. These models are trained to identify and localize objects within an image, enabling the representation of the input image as a set of feature maps extracted for each object [89, 34]. This approach acknowledges the cognitive ability of humans to rapidly interpret a scene by focusing on regions containing objects rather than processing the entire image [76].

The choice between these approaches depends on the specific dataset and the availability of annotations for the object detection task. In the case of the histopathology domain and datasets like *ARCH* [59], where obtaining features for individual objects is challenging, the images are divided into fixed-size patches, as illustrated in Figure 3.2, and fed into the Transformer. Each image patch is represented by a d-dimensional feature vector denoted as $V = \{v_i | i = 1, 2, \cdots, p, v_i \in \mathbb{R}^d\}$, where p represents the number of patches in each image.

For general-purpose datasets such as *MS-COCO* [93] and *Flickr30k* [158], where object annotations are available, a pre-trained Faster R-CNN model [121] is employed to extract 36 region features for each image [15]. In this case, 36 object features are fed into the network, while the output shape remains the same. While alternative feature extraction models like ResNet [69] or DenseNet [71] are viable, the experiments indicate that ViT is the preferred choice. The selection of ViT as the image feature extractor in this study aims to ensure consistency with the other components of the network.

3.2.2 Text Representation

Text representation is a fundamental aspect of NLP tasks, serving as a method to translate human-readable sentences into machine-interpretable vectors. The recent advent of transformer models has significantly advanced the field, consistently outperforming prior methods across a broad array of NLP tasks [96], [117], [12].

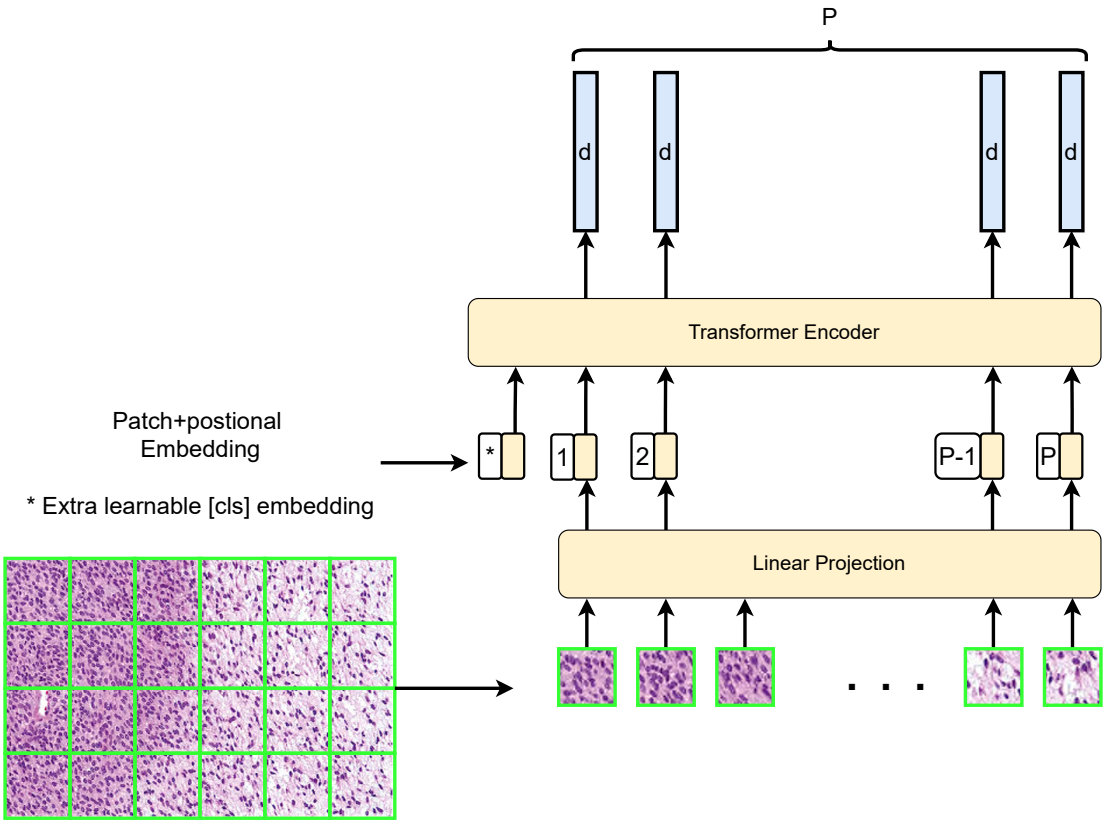


Figure 3.2: Overview of the ViT architecture: The input image is divided into a set of fixed-size patches, and each patch is linearly embedded. The total number of patches is denoted as P.

In this study, the power of the pre-trained "Roberta" architecture [96], a transformer-based model, is harnessed to encode textual input data. This model takes in a text description and outputs a feature map that embodies the semantic content of the input. The resulting text feature maps are dimensioned as (bs, seq, h), where h is the feature map dimension, seq denotes the number of tokens derived from the text, and bs stands for the batch size.

As illustrated in Figure 3.3, the Roberta model comprises two distinct stages. The first stage involves tokenizing the text data and transforming it into an embedding vector. To accomplish this, Roberta employs the WordPiece embedding [157] to tokenize the input text and map these tokens to their corresponding embeddings. Thus, for a text T consisting of seq tokens (in WordPiece embedding, a word not found in the vocabulary is subdivided

into multiple tokens), each token is embedded as a vector of size e, yielding an input size of (bs, seq, e).

The second stage involves feeding these embeddings into an encoder, which returns a sophisticated representation for each token in the context of all other tokens. Hence, a text representation is obtained for the text data T, designated as $S = s_j | j = 1, \ldots, s, s_j \in \mathbb{R}^d$. Here, each s_j encapsulates the information of each token in relation to the other tokens within the text data. Crucially, each s_j shares the same representation dimensions as v_i as discussed in section 4.2.1. This uniformity of representation dimensions facilitates the effective integration of visual and textual data in subsequent processing stages.

3.3 Multi-Head Self-Attention

Once the feature representations have been extracted for each instance within the modalities, a multi-head self-attention module is introduced to obtain m enhanced feature maps for extracted features from the previous stage. These m representations are achieved by selectively concentrating on the most pertinent pieces of information from each modality, thereby enabling the extraction of more nuanced and meaningful feature maps. The core of this module consists of a two-layer fully connected network as initially proposed in [131].

Given the input representation, denoted as $\Psi(x) \in \mathbb{R}^{seq \times d}$, the self-attention module generates m output attention feature maps, termed $\alpha_{\text{self-attention}} \in \mathbb{R}^{m \times Seq}$. The calculation for these attention maps is defined as

$$\alpha_{\text{self-attention}} = \mathbf{softmax}(w_2 \times \mathbf{tanh}(w_1(\psi(x))^T)), \tag{3.1}$$

In this equation, w_1 and w_2 are learnable parameters. These parameters are optimized during the training process to maximize the effectiveness of the attention maps in highlighting the most salient features of the input data.

Having generated the attention maps, these are then applied to the extracted features to yield the attention output, defined as $X \in \mathbb{R}^{m \times d}$. The attention output is computed as

$$X = \sigma((\alpha_{\text{self-attention}}\psi(x))w_3 + b_3), \tag{3.2}$$

In this equation, w_3 and b_3 denote additional learnable parameters, while σ represents the sigmoid activation function applied to the output. This function introduces non-linearity into the model, enhancing its capability to learn complex patterns within the data.

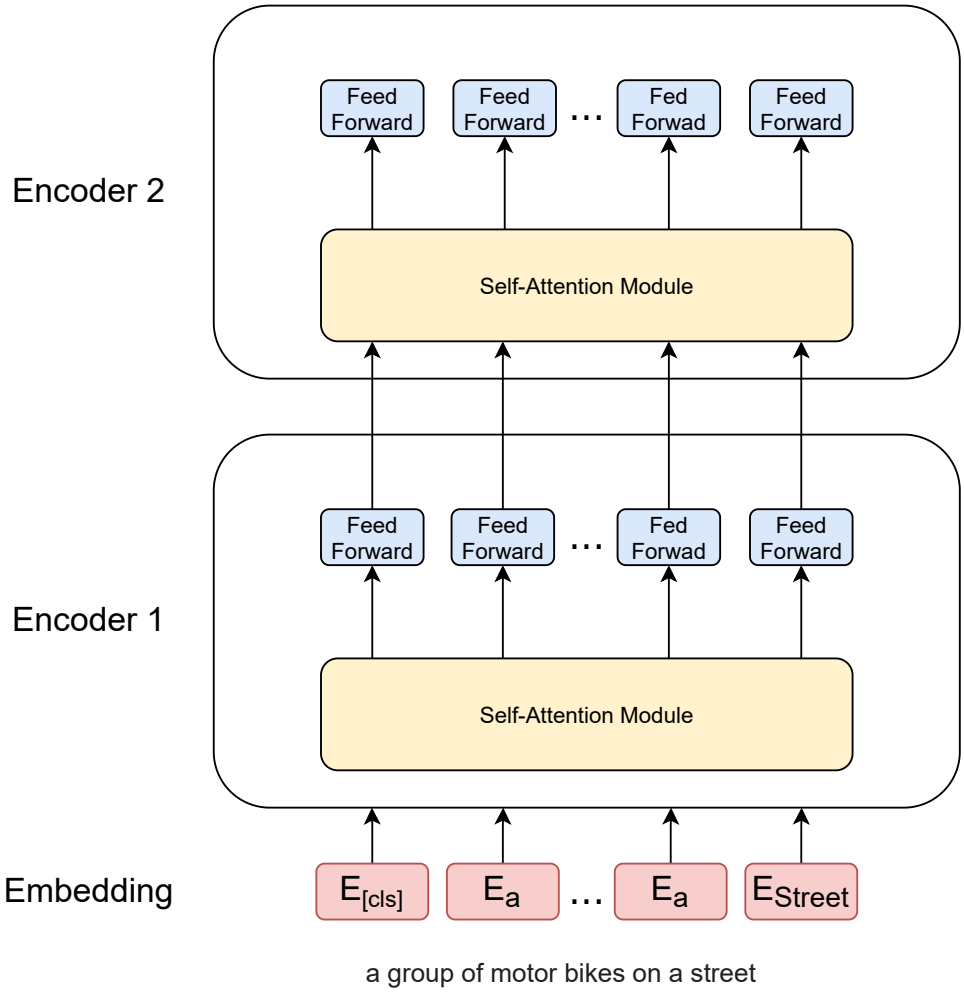


Figure 3.3: The Roberta architecture leverages tokenization to break down the input text data into individual tokens and assigns each token its corresponding embedding vector, denoted as E. The resulting embedding vectors are then passed through the encoding part of the model. The encoder is comprised of a stack of Transformer encoder blocks, with each block consisting of a self-attention module followed by a feed-forward layer. It is important to note that in this figure, only two encoder blocks are depicted, representing a subset of the complete encoder stack.

Through the implementation of this multi-head self-attention module, the model is empowered to focus on different facets of the input data, thereby capturing the intricate

relationships that exist between the features within each modality. By iteratively learning the attention weights, the module is able to emphasize the most significant information from the input, resulting in the generation of feature maps that are richer, more informative, and more discriminatory. These enhanced feature maps subsequently feed into the later stages of the model, providing a more solid foundation for the recognition of relationships between images and text.

3.4 Cross-Attention and Memory Block

Cross-modality retrieval presents a unique challenge in that the isolated examination of each modality may not yield sufficient insight into the task at hand. For effective mapping, the information from one modality must be integrated with that of the other. To tackle this issue, a cross-attention mechanism is incorporated, allowing for selective attention to distinct elements within one modality, influenced by the context derived from another [131], [72], [89].

The specific application of the cross-attention mechanism in the proposed approach is to attend to certain regions within the image, guided by the tokens present in the text input, and vice versa. The goal is to identify the relevant information in one modality that corresponds to the context provided by the other. This allows for a more nuanced understanding of the relationships between the modalities, leading to the creation of a more robust model.

The memory block functions in tandem with the cross-attention mechanism, aiming to refine the features extracted from each modality by incorporating information from the cross-attention feature maps [34]. This block serves as an enhancement to the cross-attention mechanism, further emphasizing the significant information in one modality that is relevant to the other.

The input modalities are represented as $X = \{x_i | i \in [1, m], x_i \in \mathbb{R}^d\}$ and $Y = \{y_i | i \in [1, m], y_i \in \mathbb{R}^d\}$, where X and Y can denote either image or text features. Here, m signifies the number of attention heads present in the multi-head self-attention module from the preceding step.

The cross-attention module plays a critical role in highlighting the information in modality X that is relevant to modality Y. Achieving this goal necessitates an appropriate measure to quantify the similarity between each feature map in one modality and the corresponding feature maps in the other modality. This measure forms a cornerstone

of the cross-attention mechanism, ensuring that the attention is properly focused on the most relevant aspects of the data.

The proposed solution leverages the concept of cosine similarity, a common measure extensively utilized in the field to assess the similarity between distinct feature maps [76], [34], [131], [77], [89]. Cosine similarity provides an effective means of quantifying the degree of similarity between vectors and is especially suited for high-dimensional spaces like those encountered in the context of feature maps.

In the application of the proposed solution, cosine similarity is used to ascertain the similarity between each instance in modalities X and Y, as expressed in the equation:

$$s_{ij} = \frac{x_i^T y_j}{||x_i|| \times ||y_j||}, \forall i \in [1, m], \forall j \in [1, m] \tag{3.3}$$

Here, s_{ij} denotes the similarity between the i-th feature map for modality X and the j-th feature map for modality Y. This operation serves as a core element in the cross-attention mechanism, as it establishes the correlation between the feature maps across the two modalities.

Furthermore, it is advantageous to apply a threshold operation to the similarity measure and then normalize it, as it helps eliminate the influence of negative similarity scores and ensures the magnitudes of the similarity scores remain within a manageable range [80], [89]. This operation is captured in the following equation:

$$\bar{s}_{ij} = \frac{\max(0, s_{ij})}{\sqrt{\sum_{i=1}^{m} \max(s_{ij})^2}}. \tag{3.4}$$

By using this normalization operation, the similarity scores are effectively scaled, which can assist in improving the stability of the learning process and potentially enhance the performance of the model.

In order to facilitate attention on the set Y, considering a given feature x_i in X, a weighted combination of y_j is defined. The attention function depicted in Eq. 3.4 and 3.6 represents a variation of the commonly employed "dot product attention", as detailed in numerous studies [103]:

$$a_i^x = \sum_{j=1}^{m} \alpha_{ij} y_j, \tag{3.5}$$

where

$$\alpha_{ij} = \frac{\exp(\lambda \bar{s}_{ij})}{\sum_{j=1}^{m} \exp(\lambda \bar{s}_{ij})}. \tag{3.6}$$

In Eq. 3.6, λ is the inverse temperature parameter of the softmax function [44]. This parameter allows for tuning the smoothness of the attention function, thereby providing a degree of control over the model's attention distribution.

A^x is then defined as $\{a_i^x \in [1, m], a_i^x \in \mathbb{R}^d\}$, where each element of A^x is encapsulated by the significant parts of each x_i considering the entirety of the set Y as context.

In order to further refine the extracted features of X in accordance with the salient parts of each x_i given Y, a memory unit is employed. This unit dynamically updates and refines the feature maps of X by examining both A^x and X:

$$x^* = f(x_i, a_i^x), \tag{3.7}$$

where $f(\cdot)$ can be defined in various ways [78], [152], [143], [29]. In this study, a cross-attention module for $f(\cdot)$ is adopted. As mentioned earlier, a_i^x represents the cross-attention output, highlighting the most relevant aspects of one modality in relation to the other. Meanwhile, x_i denotes the extracted features of that modality.

$$x^* = \textbf{Self-Attention}(\textbf{Concat}[x_i, a_i^x]) \tag{3.8}$$

Here, the self-attention module serves as a gate to select the most pertinent features between the concatenated information in x_i and a_i^x. The implementation of a self-attention module in tandem with a memory not only allows the refinement of feature representations but also implicitly incorporates the intra-modality relationships between modalities X and Y. This comprehensive approach ensures the model is aware of the context of each modality, thereby enhancing the performance of cross-modal retrieval.

3.5 Iterative Matching Strategy

As previously elaborated in section 3.4, memory plays a pivotal role in assisting the model to refine the feature representation, taking into account the shared information between two modalities. Employing memory within an iterative scheme can offer substantial benefits.

Each iterative step facilitates the recalibration of the X feature representation using the A^x feature map. This iterative procedure can be encapsulated as per Eq. 3.9, where k signifies the iteration step which refines the alignment for the subsequent iteration:

$$X_k^* = \textbf{Memory}(X_{k-1}, A^x). \tag{3.9}$$

For each iteration step k, the similarity score between an image (I) and text (T) is computed as follows:

$$\begin{aligned} S_k(I,T) = \alpha(\frac{1}{m}\sum_{i=1}^{m} S_k^{(v,v\rightarrow T)}(v_i,T) + \frac{1}{m}\sum_{i=1}^{m} S_k^{(w,w\rightarrow I)}(I,w_i))+ \\ (1-\alpha)(\frac{1}{m}\sum_{i=1}^{m} S_k^{(v,T)}(v_i,T) + \frac{1}{m}\sum_{i=1}^{m} S_k^{(w,I)}(I,w_i)) \end{aligned} \tag{3.10}$$

In the equation above, α is a scalar weight parameter that moderates the influence of the similarity score components. $S^{(v,v\rightarrow T)}(v_i,T)$ and $S^{(w,w\rightarrow I)}(I,w_i)$ are defined as the similarity score between image regions and text T, and text tokens and image I, respectively. These similarity scores are calculated as

$$\begin{aligned} S_k^{(v,v\rightarrow T)}(v_i,T) &= \textbf{cosine similarity}(v_i, A_k^v), \\ S_k^{(w,w\rightarrow I)}(I,w_i) &= \textbf{cosine similarity}(A_k^t, w_i). \end{aligned} \tag{3.11}$$

Further enhancements in the similarity score can be achieved by incorporating $S^{(v,T)}(v_i,T)$ and $S^{(w,I)}(I,w_i)$. These two terms are expressed in

$$\begin{aligned} S_k^{(v,T)}(v_i,T) &= \textbf{cosine similarity}(v_i, T), \\ S_k^{(w,I)}(I,w_i) &= \textbf{cosine similarity}(I, w_i). \end{aligned} \tag{3.12}$$

This inclusion allows the model to maintain the semantic meaning of each instance while concurrently striving to minimize the distance between paired instances. To consolidate all k steps, the final similarity score between image I and text T is derived as

$$S(I,T) = \sum_{k=1}^{K} S_k(I,T). \tag{3.13}$$

Here k is the number of matching steps that will be set as a hyper-parameter.

3.6 Employment of the Triplet-Loss Function

The triplet-loss function is employed, aiming to draw paired image-text data closer within the shared space while distancing them from unpaired data. This strategy has been widely recognized and applied in image-text matching domains, with numerous studies demonstrating its effectiveness [79], [85], [129], [57]. In the proposed methodology, the N-pairs bi-directional triplet-loss is incorporated, as depicted in Eq. 3.13. The principle of N-pairs loss [130], which advocates for the consideration of all negative samples within a mini-batch, rather than using a single hard negative sample [76], [89], [34], is utilized. The operation of the N-pairs triplet-loss function is visually represented in Figure 3.4.

The total loss is calculated as follows:

$$\mathcal{L}^{\mathcal{NP}} = \sum_{y_j \in mini-batch, j \neq i} \mathcal{L}(I_i, I_j, T_i, T_j), \tag{3.14}$$

$$\tag{3.15}$$

In this equation, $\mathcal{L}$ is derived from:

$$\mathcal{L} = [\Delta - S(I_i,T_i) + S(I_i,T_j)]_+ + [\Delta - S(I_i,T_i) + S(I_j,T_i)]_+. \tag{3.16}$$

In Eq. 3.16, Δ is denoted as a margin value, and $[x]_+ = max(0,x)$. $S(I,T)$, defined by Eq. 3.13, measures the similarity between an image I and a text T. This similarity score constructs a symmetric, square similarity score matrix S of size $n \times n$. During the training phase, n is defined as the size of the mini-batch. The similarity matrix's structure is demonstrated in Figure 3.5. Paired instances, images, and text with the same subscript are represented on the matrix diagonal, which should possess the highest values compared to other indices. An end-to-end training process is conducted, utilizing an optimizer, such as Adam [84].

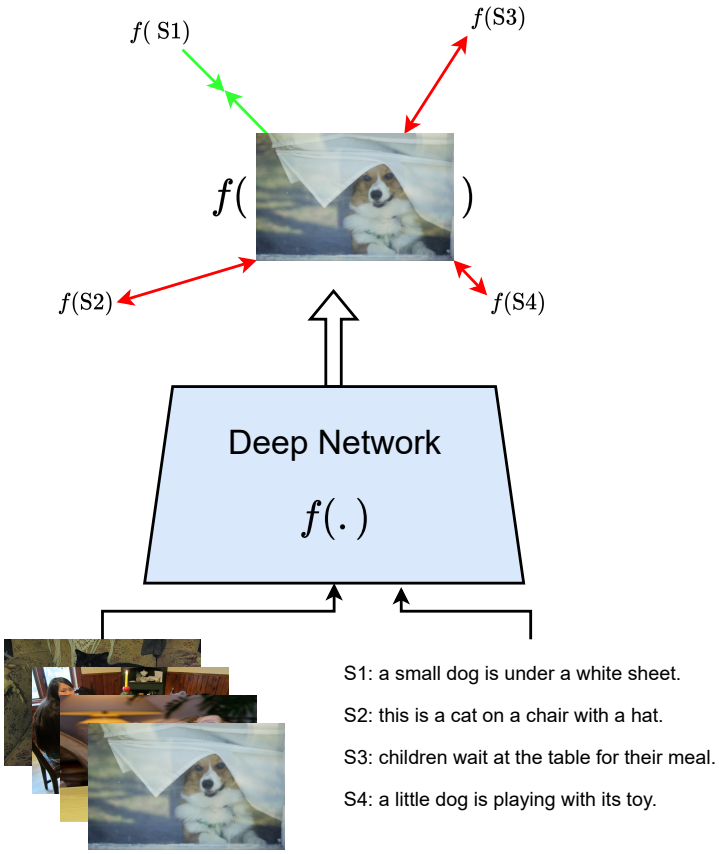


Figure 3.4: Representation of the N-pairs triplet-loss function. The arrows denote the enforced push and pull actions of the loss function. Pull connections are indicated in green, while push connections are marked in red.

3.7 Empirical Evaluation and Result Analysis

This chapter is dedicated to discussing the quantitative and qualitative findings from the cross-modality retrieval experiments conducted on various datasets. The benchmark datasets utilized in the experiments are elaborated upon in Section 3.8. Section 3.9 details the implementation strategies employed in evaluating the model. Lastly, Section 3.9.1 discusses the comparative study of the proposed method against other cross-modality retrieval techniques, with an emphasis on the superiority of the suggested approach.

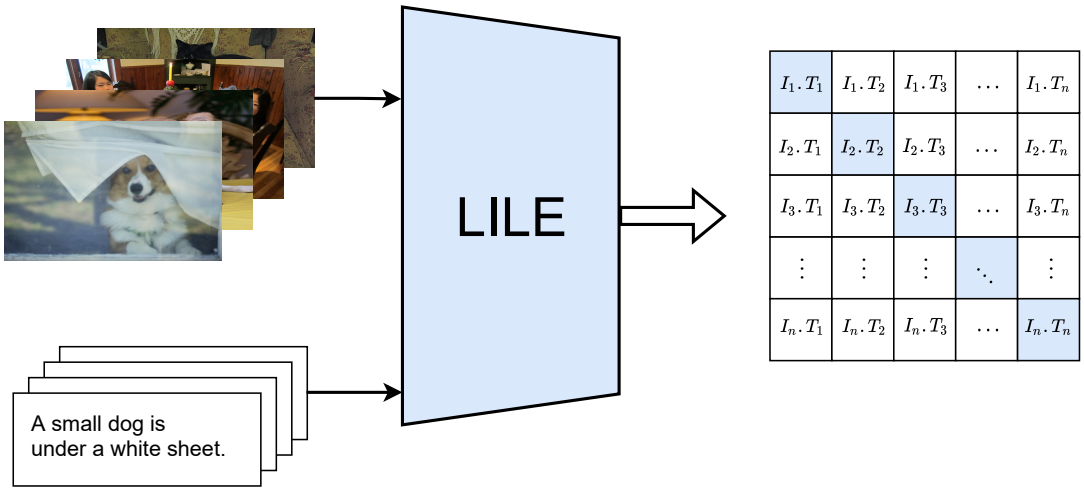


Figure 3.5: Depiction of the similarity matrix for a mini-batch with a size of n. The element $I_i.T._j$ in the matrix reveals the similarity score between the image I_i and the text T_j. Since the paired data have the same subscript, the similarity matrix is inherently diagonal.

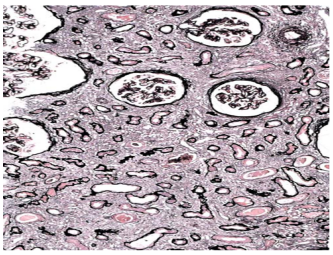
3.8 Benchmark Datasets and Evaluation Criteria

3.8.1 Benchmark Datasets

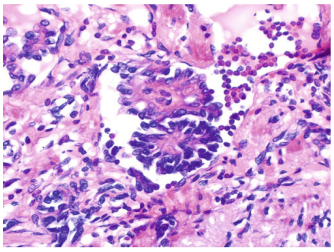
To validate the efficiency of the proposed method, two widely recognized benchmark datasets were employed in the study. They include:

1. **MS-COCO** [93]: This is a comprehensive, all-purpose dataset that includes 123,287 images along with their corresponding descriptions. Each image in the dataset is supplemented with five textual descriptions. The dataset is partitioned as follows: 113,287 images form the training set, while 5,000 images each comprise the validation and test sets. This partitioning is consistent with the methodologies adopted in previous studies [79], [89], [57]. The performance assessment was carried out by averaging the results obtained over five sets of 1,000 test images and by testing the entire set of 5,000 test images, following the procedures outlined in [89],[34].

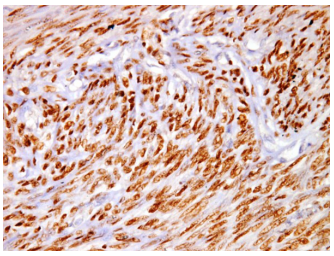
2. **ARCH** [59]: Serving as a critical dataset for CP, ARCH is a multiple-instance captioning collection that extensively covers a broad spectrum of tissue types and staining. This



Cortex in a renal biopsy specimen from a man of 38 known to have membranous nephropathy, shown by a biopsy nearly 2 years before this one. The nephrotic syndrome had persisted since the first biopsy, and acute renal failure had developed 1 month before this biopsy. There is now widespread uniform tubular atrophy, which suggests longstanding renal vein thrombosis. This was confirmed by radiologic investigation. Renal function did not recover



Papillary intralymphatic angioendothelioma. This is a dermal or subcutaneous lesion of children or adults, with intravascular growth of cells that have a lymphatic endothelial immunophenotype. Some have adjacent lymphangiomas or clusters of lymphatic vessels.The earlier terminology was malignant endovascular papillary angioendothelioma of childhood , but in a more recent series, reported cases did not recur or metastasize



High PR expression (3+) in a leiomyoma,(PR x400)

Figure 3.6: A selection of three diverse samples from the ARCH dataset. These images reflect the dataset's broad range of staining techniques and normalization processes. Alongside each image, its associated description showcases the varied structure and depth of narrative detail that characterizes the dataset. This diversity not only speaks to the robustness of the dataset but also highlights the complexity of the task at hand - aligning visual and textual modalities that span a wide array of contexts and styles.

dataset is enriched with detailed morphological descriptions and diagnoses, providing a comprehensive outlook on the subject matter. Figure 3.6 offers a glimpse of the sample data contained in the ARCH dataset. The images and their corresponding annotations in this dataset were meticulously sourced from an array of PubMed medical articles and authoritative pathology textbooks. Comprising 7,579 images, each accompanied by a descriptive annotation, the ARCH dataset presents a substantial amount of data for deep analysis. In the context of the experiments conducted for this study, a 5-fold cross-validation method was employed to ensure a rigorous evaluation of the model's performance on this dataset.

3.8.2 Evaluation Metric

The effectiveness of the proposed method is evaluated using the Recall at K (R@K) metric, a standard measure widely employed in the realm of information retrieval [34, 89, 113, 102]. This metric quantifies the proportion of queries that are correctly retrieved within the top K search results, providing a measure of the model's precision and relevance.

For the task at hand, the R@k metric is applied to both directions of cross-modal retrieval: text-to-image and image-to-text. Specifically, the model is assessed on its ability to

retrieve relevant images given a textual query (Text-to-Image), and conversely, to identify pertinent textual descriptions in response to an image input (Image-to-Text).

A comprehensive evaluation is undertaken using three different thresholds for K: R@1, R@5, and R@10. This approach ensures the model's performance is analyzed in terms of its ability to retrieve the correct match at varying levels of precision. Additionally, to provide a broader perspective on the model's overall efficiency, a cumulative measure termed "R@sum" is also reported. This metric, as employed in prior studies [72], is the sum of the recall rates at the different K thresholds, i.e., R@1+R@5+R@10. Thus, the "R@sum" provides a holistic performance view across multiple levels of precision.

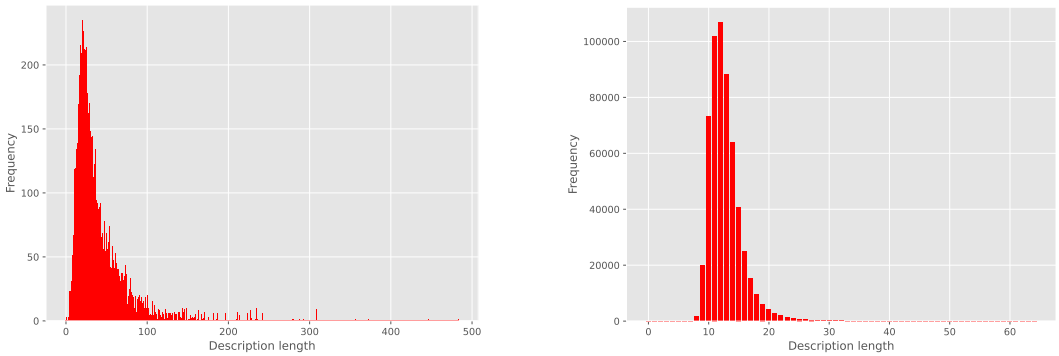
3.9 Implementation Details

The execution of the experiments employing the ARCH dataset utilized the DeiT (Data-efficient Image Transformers) architecture [139] as the primary image encoder. This architecture, comprising 12 layers and 768 hidden states, decomposes the input image into patches of 16×16. As a result, each processed image effectively translates into a grid of 196 patches, each with dimensions 16×16.

Prior to its integration into the main model, the DeiT-based image encoder was individually trained for a classification task using the dataset featured in the "KimiaNet" paper [122]. This preliminary training phase facilitated the initialization of the weights for the image encoder in the subsequent stages of the main experiment. Due to the limited size of the dataset, only the final two blocks of the image encoder were set as trainable, while the remaining sections of the network were frozen to prevent over-fitting [149].

For the text encoder, the weights were initialized using the BioMed-RoBERTa model [66]. This model, trained on a vast corpus of 2.68 million scientific papers from the Semantic Scholar database [10], is a language model predicated on the RoBERTa-base architecture [96]. It comprises 12 layers, 768 hidden dimensions, and accommodates 110M parameters. Mirroring the approach taken with the image encoder, only the last two layers of the text encoder were designated as trainable, with the rest of the network frozen.

The ARCH dataset presented a unique challenge due to the variability in the lengths of the descriptions accompanying the images, as depicted in Figure 3.7. These descriptions ranged significantly in length, from as few as 2 tokens to as many as 484 tokens. This non-uniformity, coupled with the fact that the descriptions were generally longer than those in the MS-COCO dataset, necessitated a tailored solution. To address these issues, each description was segmented into individual sentences. Subsequently, new descriptions



(a) (b)

Figure 3.7: (a) Frequency distribution of the token lengths in the descriptions of the ARCH dataset, showcasing the variability in text lengths. (b) Frequency distribution of the caption lengths in the MS-COCO dataset, highlighting the range of text lengths. The MS-COCO dataset exhibits an average of 13.81 tokens per caption, with a maximum of 65 tokens, while the ARCH dataset demonstrates an average of 51.49 tokens per description, with a maximum of 484 tokens. These statistics emphasize the substantial differences in text length between the two datasets.

were generated by concatenating the first sentence with every subsequent sentence. This approach effectively enabled each image to be associated with multiple descriptions, thereby augmenting the dataset and promoting a more uniform distribution of description lengths.

For the experiments utilizing the MS-COCO dataset, the same architectures were implemented for the image and text encoders as those employed in the ARCH dataset experiments. However, certain modifications were made concerning the weight initialization process and the number of trainable layers.

The weights for the image encoder were initialized from a model pre-trained on the ImageNet-1k dataset [49], a comprehensive dataset comprising one million images categorized into 1,000 classes. The input data for the image encoder was the feature maps that had been extracted using the Faster R-CNN algorithm, as detailed in section 4.2.1. In this case, the last six layers of the image encoder were rendered trainable, with the remaining layers set as frozen. For the text encoder, the weights were initialized from a pre-trained Roberta-base network, which was trained on a comprehensive compilation of five general datasets[1]. Similar to the image encoder, the final six layers of the text encoder were set to

[1]This includes BookCorpus [6], which comprises 11,038 unpublished books; English Wikipedia [1]; CC-

be trainable, while the rest were frozen.

The experiments were conducted using the PyTorch V1.3 framework, with computations facilitated by two NVIDIA V100 GPUs. Prior to their introduction into the model, images were resized to 224×224 pixels and normalized. The hyper-parameter m, which denotes the number of self-attention heads, was optimally set at 16 for all experiments. The best performance was achieved by setting α to 0.8 in the loss function during the experiments.

The model was trained using the Adam optimizer [84] with an initial learning rate of $1e-4$. In order to optimize model performance, the learning rate was adjusted downward whenever the evaluation metric ceased to improve. This dynamic approach to learning rate adjustment ensured that the model continually evolved toward optimal performance.

3.9.1 Experiments and Results

The performance of the method introduced in this study, referred to as LILE, is benchmarked against a selection of leading-edge methods using the two datasets outlined in Section 3.8. For the ARCH dataset, it is important to note that models that are not publicly available were excluded from this comparison.

The findings are shown in Tables 3.1 and 3.2 for the MS-COCO and ARCH datasets, respectively. The results indicate that the proposed deep learning method, LILE, demonstrates superiority over the preceding top-performing model, IMRAM [34]. The overall performance of LILE under the R@sum metric surpasses that of IMRAM by a margin of 6.3 and 8.3 in the MS-COCO(1K) and MS-COCO(5K) datasets, respectively.

When examining the recall metrics for text retrieval at different K values, LILE achieved a maximal performance improvement of 1.0 and 1.9 for R@1 in the MS-COCO(1K) and MS-COCO(5K) datasets, respectively. In terms of the image retrieval task, the highest performance improvements observed were 2.4 for R@1 in both MS-COCO(1K) and MS-COCO(5K).

Since no results have been previously reported on the ARCH dataset, publicly available methods such as IMRAM [34], PVSE [131], and CLIP [116] were selected and adapted to evaluate their performance on this dataset. The proposed framework, LILE, outperformed the best previous model, IMRAM, by a significant margin of 28.1 in terms of overall performance under the R@sum metric.

News [7], which includes 63M English news articles; OpenWebText [5], a web text dataset used to train GPT-2; and Stories, a subset of the CommonCrawl data [8].

Table 3.1: Comparison with state-of-the-art methods on MS-COCO

Method	Text Retrieval			Image Retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
1K							
DPC [168]	65.6	89.8	98.5	47.1	79.9	90.0	467.9
SCO[73]	69.9	92.9	97.5	56.7	87.5	94.8	499.3
SCAN[89]	72.7	94.8	98.4	58.8	88.4	94.8	507.9
PVSE[131]	69.2	91.6	96.6	55.2	86.5	93.7	492.8
VSRN[90]	76.2	94.8	98.2	62.8	89.7	95.1	516.8
IMRAM[34]	76.7	**95.6**	**98.5**	61.7	89.1	95.0	516.6
OSCAR (Fine-tuned)[91]	89.8	98.8	99.7	78.2	95.8	98.3	563.6
LILE	**77.7**	95.4	98.3	**64.1**	**91.0**	**96.4**	**522.9**
5K							
DPC [168]	41.2	70.5	81.1	25.3	53.4	66.4	337.9
SCO[73]	42.8	72.3	83.0	33.1	62.9	75.5	369.6
SCAN[89]	50.4	82.2	90.0	38.6	69.3	80.4	410.9
PVSE[131]	45.2	74.3	84.5	32.4	63.0	75.0	374.4
VSRN[90]	53.0	81.1	89.4	40.5	70.6	81.1	415.7
IMRAM[34]	53.7	**83.2**	91.0	39.7	69.1	79.8	416.5
ALIGN (Zero-shot)[77]	58.6	83.0	89.7	45.6	69.8	78.6	425.3
ALIGN (Fine-tuned) [77]	77.0	93.5	96.9	59.9	83.3	89.8	500.4
OSCAR (Fine-tuned)[91]	73.5	92.2	96.0	57.5	82.8	89.8	491.8
CLIP[116]	58.4	81.5	88.1	37.8	62.4	72.2	400.4
LILE	**55.6**	82.4	**91.0**	**41.5**	**72.1**	**82.2**	**424.8**

These results are indicative of the efficacy of the proposed approach in handling cross-modality retrieval tasks, further solidifying its position as a SOTA solution in this domain. The analysis herein highlights the potential of LILE to shape future progress in the area of cross-modality retrieval.

The proposed approach, when trained on the MS-COCO dataset, surpasses all alternative methodologies in terms of prevalent benchmarks in the field, often by a considerable margin (as detailed in Table 3.1). This underscores that the innovative architecture and

objectives outlined in this proposal significantly boost the efficacy of the retrieval task.

As discussed in Chapter 2, cross-modal retrieval tasks can gain substantial benefits from the deployment of voluminous data and robust computational resources. Hence, methods trained on large-scale datasets (such as CLIP [117], ALIGN [77], and OSCAR [91]) cannot be directly compared with those trained exclusively on smaller datasets like MS-COCO (including DPC [168], SCO [73], SCAN [89], PVSE [131], VSRN [90], IMRAM [34], and LILE). The values associated with methods trained on substantial datasets are reported primarily to underscore this distinction.

The most analogous comparison to the proposed approach lies with the IMRAM method [34]. Both methodologies share the concept of applying an iterative matching scheme through a memory network to refine image and text feature representations. However, in LILE, feature representations rely on a multi-head self-attention module. This module enables the model to selectively focus on the most relevant parts of each instance. Furthermore, the additional term incorporated into the loss function provides the model with a more explicit source of information to align paired image and text instances.

The results obtained from the ARCH dataset demonstrate an even more significant improvement compared to the MS-COCO results. One plausible reason for this enhancement could be the application of a Transformer for encoding the input data. As detailed in Section 3.9, a Roberta-based architecture is employed for encoding descriptions. This model, trained on a large quantity of scientific text, aids in learning a more robust representation of medical descriptions.

In contrast, IMRAM [34] and PVSE [131] both implement a Gated Recurrent Unit (GRU) [42] in their networks to encode text data. When descriptions are lengthy, as is often the case with medical reports and ARCH datasets, their text encoders may struggle to encode the descriptions appropriately. As mentioned in Chapter 2, when the length of the input sequence grows excessively long, the ability to retain context information plummets significantly [56]. Consequently, the application of a Transformer-based model for the text encoder can significantly enhance the network's performance.

Furthermore, the analysis of results on the ARCH dataset provides compelling evidence that methods like CLIP [116], trained on vast datasets, may not perform satisfactorily on specialized tasks such as image-text retrieval in pathology. This particular type of task seemingly requires more sophisticated methodologies, such as the framework proposed in this study.

Exploring the Impact of the Iterative Matching Scheme, K. To gain a more comprehensive understanding of K's influence within the proposed methodology, K is incrementally increased from 1 to 4. The model is subsequently trained and tested on the

Table 3.2: Comparison with previous methods on ARCH.

Method	Text Retrieval			Image Retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
PVSE	36.2	48.8	53.3	29.3	43.8	52.6	264.0
IMRAM	40.3	52.5	59.5	32.4	48.2	56.1	289.0
CLIP (Zero-shot)	38.2	51.1	57.0	30.8	46.1	53.2	276.4
LILE	**44.8**	**57.4**	**64.4**	**36.7**	**52.2**	**61.7**	**317.1**

Table 3.3: The effect of the number of matching steps, K, in MS-COCO dataset.

K	Text Retrieval			Image Retrieval			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
1	77.2	96.4	99.0	64.4	92.0	96.6	525.6
2	79.6	97.4	99.7	67.0	92.8	97.2	533.7
3	79.9	97.0	99.0	66.2	92.3	97.3	531.7
4	78.5	96.7	98.8	65.9	92.5	97.4	529.8

MS-COCO (5K) benchmark dataset. Table 3.3 elucidates the results for R@1, R@5, R@10, and R@sum metrics in both image-to-text retrieval and vice versa.

Upon examining the results, it becomes evident that when $K = 2$, the model consistently outperforms its counterparts with different K values. The gap in performance between $K = 1$ - which serves as a baseline without the iterative scheme - and other K values is notably significant. This disparity underscores the pivotal role of the iterative scheme in enhancing the efficacy of the proposed approach.

Chapter 4

Self-Supervised Learning

4.1 Introduction

As stated in Chapter 1, the primary objective of this thesis is to develop a novel methodology for image-text cross-modal retrieval that efficiently captures the inherent relationships between images and their corresponding text. Additionally, this thesis aims to tackle the challenge of scarcity in labelled data, particularly in the case of pathology images, through a SSL approach.

This thesis introduced a novel patching scheme tailored specifically for processing WSI in histopathology. This scheme is intended to adapt to the intrinsic features of WSIs, thereby assisting the model in generating more robust and meaningful embeddings for the visual component. These objectives aim to make a significant contribution to the field of Artificial Intelligence (AI) by advancing SOTA methods in image-text cross-modal retrieval, especially in domains with limited availability of large-scale labelled data.

The proposed methodology encompasses several key components that work collectively toward accomplishing the objectives. The forthcoming sections will delve into the intricate aspects of these components, providing a comprehensive understanding of their design and functionality.

The proposed approach relies on the foundation that has been discussed in Chapter 3. The suggested method aims to capture the most salient parts of the image and text as described in Section 3.2. In this context, a SSL method is applied to have a robust feature representation for the given image. In Section 4.2.1, a thorough exploration of the proposed SSL approach and its incorporation into the suggested method are provided.

Section 3.6 provides an in-depth explanation of the proposed loss function. This loss function is meticulously designed to align with the objectives of this study, providing a coherent means of evaluating and driving the learning of the model. By combining these components, the proposed method aims to advance the SOTA in image-text cross-modal retrieval tasks for domains with limited access to large-scale (gigapixel) labelled data.

4.2 Feature Representation

Data representation, also known as representation learning, is of paramount importance in the sphere of ML algorithms. Its significance lies in its ability to empower models to independently unearth the crucial representations embedded in the supplied data 3.2.

Utilizing SSL methods for feature representation can provide substantial benefits, particularly when dealing with a scarcity of labelled data [164], [110]. These models have the capability to learn intricate patterns and meaningful representations directly from the unlabelled data itself. They achieve this by defining and solving *pretext tasks* that implicitly require understanding the structure of the data, thereby learning useful features without the need for explicit labels.

The resulting representations are not only meaningful but also robust. This strength stems from the fact that the representations are trained on a larger portion of the available data, not just the subset that is labelled. This robustness, coupled with expressive data representation, greatly enhances the ability of models to capture the latent structure of patterns in the data. Consequently, the performance of the models is bolstered, and their capacity to generalize across varying contexts is significantly amplified. This inherent ability to generate insightful features from unlabelled data makes SSL models an especially potent tool in scenarios where acquiring labelled data is challenging, costly, or time-consuming [18]. In the domain of medical imaging, obtaining a large amount of accurately labelled data is commonly considered a significant hurdle.

SSL models are designed to mitigate this problem, leveraging the wealth of unlabelled data to learn transferable feature representations. Consequently, these models can help to bridge the gap between the data-rich and data-poor settings, advancing performance in tasks such as image-text cross-modal retrieval. They do so by training on the unlabeled data to learn rich feature representations that capture the underlying data distributions.

In this study, the challenge of sparse annotated data is addressed, particularly noticeable in the visual domain. A SSL approach combined with Transformer architecture is deployed for the extraction of image features. The unique features of patches in WSIs

render the fusion of Transformer architecture and an SSL approach remarkably fitting for extracting meaningful image features. This approach helps capitalize on the vast amount of unlabelled image data, enabling the extraction of pertinent features while reducing the dependency on labelled data. Additionally, the implementation of an SSL scheme for the visual component promotes enhanced generalization to unseen data to generate a more robust feature representation for the given image.

Contrarily, when it comes to the textual modality, the dearth of labelled data is generally not a pressing concern. Thus, Transformer architecture, devoid of SSL, is utilized for the extraction of textual features.

Through the utilization of Transformers' powerful representation learning capabilities and the implementation of SSL for image feature extraction, this research aims at capturing and exploiting the intricate interrelationships between visual and textual information.

4.2.1 Image Representation

The main concept of this study relies on the concept that the model applied a SSL to extract image features. In this thesis, the concept that has been used in the DINO method [32] has been modified with a novel patching approach that is specialized for WSIs. The detail of this patching scheme is elaborated in Section 4.2.2. The DINO framework follows a similar overall structure to recent self-supervised approaches [31], [40], [37]. However, the method also utilizes the concept of *knowledge distillation*. Knowledge distillation is a learning technique in which a student network, denoted g_{θ_s}, is trained to imitate the output of a pre-existing teacher network, denoted g_{θ_t} [70]. For a given input image x, the student network with parameters s learns to generate probability distributions, P_s, over k dimensions. In a similar fashion, the teacher network with parameters t also generates probability distributions, P_t. These probabilities, denoted by P, are derived by applying a softmax function to the network's outputs and normalizing them. More precisely, each probability distribution can be calculated by Eq. 4.1. In this equation for the student and teacher networks, different τ will be applied to control the output distribution's sharpness:

$$P^{(i)} = \frac{exp(g_\theta(x)^{(i)})/\tau)}{\sum_{k=1}^{K} exp(g_\theta(x)^{(k)}/\tau)}, \tag{4.1}$$

The process of aligning the output distributions of the teacher and student models plays a crucial role in knowledge transfer and learning. By minimizing the cross-entropy loss with respect to the parameters of the student network, as indicated in Eq. 4.2, the

model ensures that the student network learns to match the distributions produced by the teacher network. This encourages the student network to capture the same patterns and knowledge as the teacher, effectively transferring the learned information.

Minimizing the cross-entropy loss allows the model to measure the dissimilarity between the predicted probability distributions of the student network and the actual distributions from the teacher network. Through back-propagation and gradient descent optimization, the parameters of the student network are adjusted to reduce this dissimilarity, enabling the student network to approximate the behaviour of the teacher network. This alignment of distributions facilitates the transfer of knowledge, as the student network can learn from the teacher's expertise and benefit from its learned representations.

This knowledge transfer process has several advantages. First, it helps the student network to generalize and capture complex patterns and relationships present in the data, even when the labelled data is limited. By leveraging the knowledge encoded in the teacher network, the student network can learn from a more comprehensive and diverse set of examples, leading to improved performance.

Additionally, aligning the output distributions facilitates model regularization. The student network is encouraged to produce outputs that are more confident and accurate, as it aims to match the teacher's distributions. This regularization helps to prevent overfitting and enhances the generalization ability of the student network.

$$\min_{\theta_s} -P_t(x)log(P_s(x)) \tag{4.2}$$

The integration of knowledge distillation within the DINO framework has been suggested by its authors, involving the creation of various crops from a single input image [31]. This led to the generation of a set V, which contained multiple views of the input image. Specifically included within this set are two global views along with several local views, all produced at lower resolution.

Adapting to the unique requirements of WSI and the patch extraction methods used in histopathology imaging, this set V has been revised in the proposed methodology. The original DINO paper recommended using 96x96 patches for local views and 224x224 for global patches and scaling the patches as needed. However, when extracting patches from a WSI at a 20x magnification - the most common approach in digital pathology applications up-sampling or down-sampling of the local or global views could potentially alter the image's magnification, thereby introducing inconsistencies within the network.

To address these challenges, a tailored patching solution will be introduced, to seamlessly align with the inherent nature of WSI patching techniques. This enhanced approach

has been termed **Harmonized DINO** (H-DINO). Figure 4.1 shows the architecture of this block. More comprehensive details of this refined patching method can be found in Section 4.2.2.

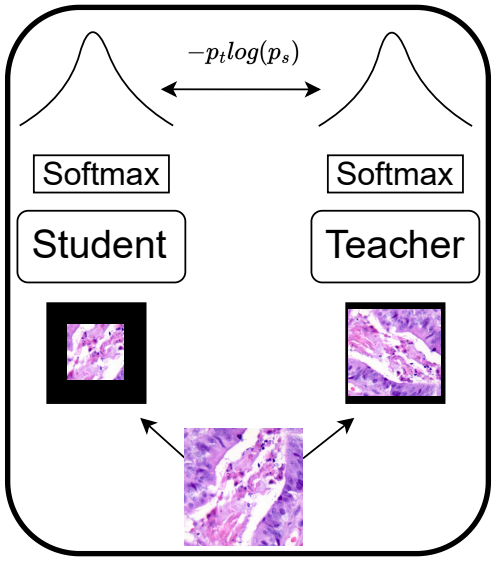


Figure 4.1: Diagram of the H-DINO Architecture: Illustrating both the student and teacher networks. For the sake of clarity, only one patch per view is represented.

4.2.2 Patching

In alignment with the discussion from the previous section, it is highlighted that the patching technique proposed is significantly more congruent with the inherent characteristics of WSIs. The process of patch extraction has been designed to seek similarities within the *neighbouring region of a specific patch* instead of looking into that specific patch and finding local and global crops as proposed in the DINO paper [32].

In pursuit of this extension to DINO, an approach has been adopted where the extraction of patches is carried out on a scale larger than the final size intended to be inputted into the network. This is accomplished by enlarging the patch size around the specific location based on the dataset and patching method. This adjustment to the patching method is more reliable compared to other augmentations, such as adding noise or blurring the data. Preserving the image's original integrity and utilizing neighbouring regions

that share context with the target area aids the model in achieving a more robust representation, in conjunction with other augmentations. The visualization of this step can be seen in Figure 4.2. Subsequently, two categories of views – more global and more local – are extracted from the expanded patch, with each being subjected to its respective set of transformations. The distinction between these views is fundamentally determined by the size of the cropped region. The cropping range for the more local views is set between 50 and 140, while for the more global views, the cropping size is set between 140 and 224. Given that uniformity in magnification is maintained in all selected patches, the core features of these patches remain unaltered. This departure from the DINO's patching approach is crucial for computational pathology. This consistency ensures that each patch holds its unique representation and characteristics intact, which in turn assists in maintaining the consistency of the information it offers.

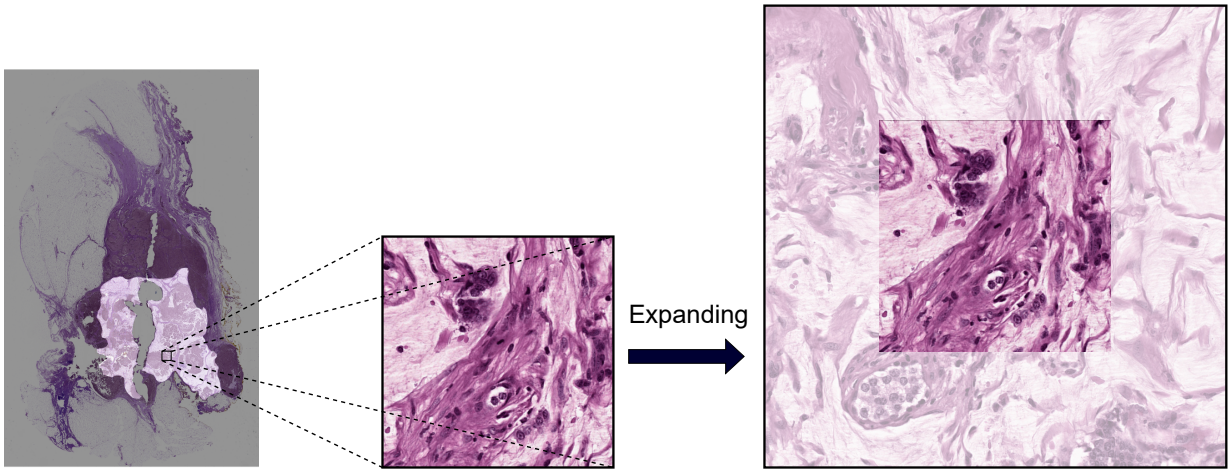


Figure 4.2: Diagrammatic Demonstration of Patch Extraction and Expansion from a WSI for Augmentation: Progressing from left to right, the first step in the process is depicted - the mask and the WSI are overlaid to highlight the region from which patches are extracted. For ease of understanding, a single extracted patch is displayed. The next phase of the procedure is demonstrated in the subsequent image, where the size of the extracted patch is expanded. This enlarged output is then fed into the final transformation stage.

Moreover, the uniformity in magnification also guides the network's learning process in a meaningful way. By processing patches that are expected to share similar attributes, the network learns to map related concepts or visual features closer to each other in its internal representation space. This capacity is expected to be central to the network's ability to distinguish between different classes and categories and, thereby, crucial to achieving high

performance on classification or retrieval tasks in the forthcoming experiments.

These transformations are carefully selected to match the specific network into which the patch will be integrated. A comprehensive visualization of this entire process can be found in Figure 4.3. This figure depicts the interplay among the extraction, enlargement, and transformation of patches, all contributing to the distinct approach that has been adopted in the design.

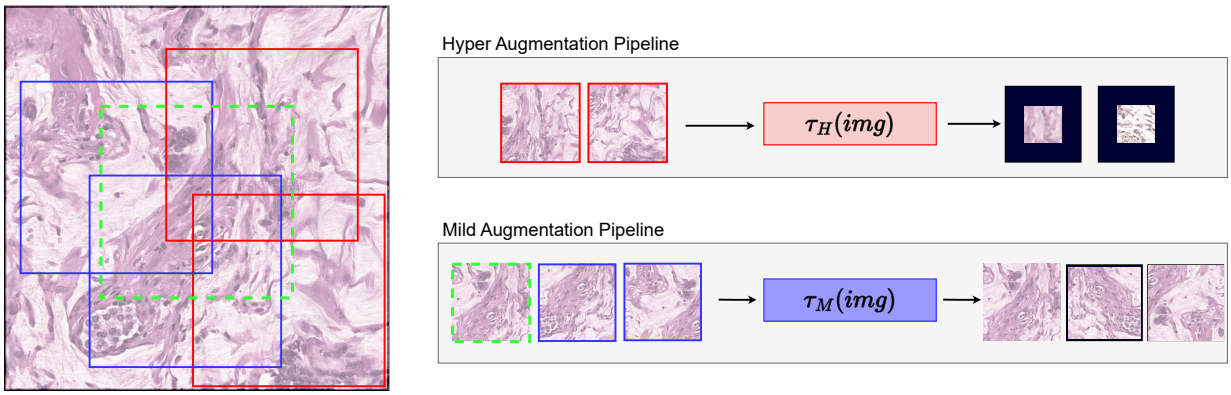


Figure 4.3: Visual Representation of the View Selection and Augmentation Pipeline: The illustration on the left depicts the process of random view selection, demonstrating the extraction of two views each for the Hyper augmentation (τ_H) and Mild augmentation (τ_M) sets. For clarity, only two instances of each set are displayed. These views are randomly cropped from an expanded patch region, each measuring 224 by 224 pixels. The diagram on the right portrays the augmentation pipeline, detailing two unique augmentation sets with varying intensity levels chosen for the Hyper augmentation and Mild augmentation pipelines. The impact of these differing intensities is further highlighted by the distinct outputs they produce.

After the patch selection, the more robust and less noisy patches are passed through the teacher model, while all the patches that have been generated in the patch selection stage are passed through the student model. Both teacher and student networks are identical in architecture with different sets of parameters θ_s and θ_t for student and teacher networks, respectively.

After extracting robust feature representations for images, the embeddings for the textual part will be extracted as described in Section 3.2.2. The implementation described in Chapter 3 is followed to match images and texts and generate the similarity matrix.

4.3 Loss Function

The proposed loss function guides the training of the entire model in an end-to-end paradigm which helps the network to have better performance and more flexibility [61]. It means that the proposed loss function should contain two parts. The first part should draw paired image-text data closer within the shared space while making the unpaired data far from each other:

$$\mathcal{L} = [\Delta - S(I_i, T_i) + S(I_i, T_j)]_+ + [\Delta - S(I_i, T_i) + S(I_j, T_i)]_+. \tag{4.3}$$

This part of the loss function has been elaborated in Section 3.6.

Meanwhile, the second part of the loss function aims to force the vision model to generate a more robust feature representation for the given images using a SSL method. In the training procedure, the loss will be minimized with respect to θ_s:

$$\min_{\theta_s} \sum_{x \in \left\{x_1^M, x_2^M\right\}} \sum_{\substack{x' \in V \\ x' \neq x}} -P_t(x) log(P_s(x')). \tag{4.4}$$

The proposed patching scheme, discussed in detail in Section 4.2.2, generates multiple views of a given image, which are assembled into a set V. This set is characterized by the inclusion of two distinct subsets of image views, differentiated by the type of augmentation applied. One subset comprises views generated through the application of "Mild Augmentation" (τ_M), while the other uses "Hyper augmentation" (τ_H).

In the context of Eq. 4.4, x is a selection made from the subset of views created through the τ_M technique. In contrast, x' is chosen from the collective set of all views, excluding the specific view selected for x. Eq. 4.4 seeks to minimize the distribution distance between these various views of the same image, which is a critical step towards achieving a more robust image understanding. This minimization process ultimately assists the model in distinguishing subtle differences within the image and understanding how different elements of the image relate to each other.

The proposed loss function here can be applied on any number of patches; however, as the authors in the DINO paper have suggested, the proposed method only used two patches for the teacher network. In the approach under discussion, the absence of an initial teacher

model, denoted by g_{θ_t}, necessitates its iterative construction from previous versions of the student network. As suggested by the authors in the DINO paper [32], an approach that remarkably aligns with this framework involves applying an exponential moving average (EMA) to the student weights and a strategy often referred to as a momentum encoder. The update rule for this strategy is described by

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s, \tag{4.5}$$

where λ undergoes a gradual transition from 0.996 to 1 during the training process, following a cosine schedule. This dynamic tuning of λ orchestrates the continuous updating of the teacher network parameters (θ_t) based on the concurrently evolving student network parameters (θ_s). This symbiotic process effectively transfers learned representations and knowledge between the teacher and student networks, promoting consistent learning progress and model robustness.

In the context of knowledge distillation models, collapse is a challenging issue that can severely limit the model's learning ability and overall performance. The term "collapse" generally refers to a situation where the network outputs become excessively homogeneous or overly focused on a single feature, thereby failing to capture the diversity and complexity inherent in the data. In extreme cases, this can cause the model to converge to a trivial solution, rendering it ineffective for its intended task.

In their work on the DINO approach, authors propose a couple of mechanisms to mitigate the risk of collapse: sharpening and centring.

Sharpening: This method modulates the temperature of the softmax function in Eq. 4.1, effectively increasing the peakiness of the teacher outputs. By making the outputs more distinct, sharpening discourages the model from collapsing towards trivial or overly generalized solutions. The temperature value is typically initialized to a high value (making the softmax outputs more uniform) and is gradually decreased during training, encouraging the development of more distinct, "sharpened" outputs.

Centering: The process of Centering, as employed in the DINO framework, dynamically fine-tunes the teacher network's outputs before applying the softmax operation. This maneuver, detailed in Eq. 4.6, effectively ensures the teacher network's outputs are centred around a common point. More specifically, it subtracts the mean feature vector, derived from all data points within a batch, from each individual data point's feature vector. This step helps prevent any single dimension from dominating and reduces the risk of model collapse. Such a centring process can be metaphorically interpreted as integrating a bias term

"c", which could be negative, into the network output, consequently shifting or offsetting the output values for better balance and stability:

$$c \leftarrow mc + (1-m)\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}(x_i), \tag{4.6}$$

where $m > 0$ determines how rapidly the center vector is updated over time, B is the batch size and g_{θ_t} refers to the teacher network. In the proposed method centring and sharpening are applied to the teacher model.

After calculating the loss functions for the self-supervised task based on Eq. 4.4 and the cross-modal task based on Eq. 4.3, the total loss is computed as follows:

$$\begin{aligned} \mathcal{L}_{total} &= \beta\mathcal{L}_{ssl} + \gamma\mathcal{L}_{cross-modal}, \\ \mathcal{L}_{ssl} &= \sum_{x\in\left\{x_1^M, x_2^M\right\}} \sum_{\substack{x'\in V \\ x'\neq x}} -P_t(x)log(P_s(x')), \\ \mathcal{L}_{cross-modal} &= [\Delta - S(I_i,T_i) + S(I_i,T_j)]_+ + [\Delta - S(I_i,T_i) + S(I_j,T_i)]_+. \end{aligned} \tag{4.7}$$

The weight parameters, denoted by scalars β and γ, play a crucial role in determining the impact of each loss term. They regulate and moderate the influence of these terms on the overall loss function. $\mathcal{L}_{ssl}$

4.4 Datasets and Experiments

This section delves into three integral components of the study. The initial portion provides a comprehensive overview of the datasets utilized for evaluating the proposed method, shedding light on their characteristics, significance, and how they effectively lend themselves to this study's objectives.

In the latter part of this section, the metrics that have been employed for conducting the experiments are expounded upon. This involves a thorough explanation of the reasons underlying the selection of these specific metrics, highlighting their importance in facilitating a precise and insightful assessment of the experimental results.

Subsequently, the focus shifts to exploring the metrics used, the design of the experiments, and a thorough analysis of the results obtained. This segment delves into the experimental setup, the reasoning behind the chosen metrics, and the rigorous evaluations conducted. Moreover, the results are dissected and interpreted, providing insightful perspectives and highlighting the implications of the findings.

4.4.1 Benchmark Datasets

In order to rigorously assess the efficacy of the proposed methodology, three distinct benchmark datasets specific to the pathology domain have been selected and utilized. Each of these datasets serves as a comprehensive representation of different aspects and challenges within the field of pathology, thus offering a robust environment to validate the suggested method.

Grand River Hospital (GRH) Breast Cancer: This collection comprises 173 WSIs, obtained from a wide-ranging patient population diagnosed with an array of breast cancer subtypes at the GRH. These primary diagnoses are classified into 22 distinct categories, marking the broad spectrum of breast cancer subtypes. A detailed breakdown of these categories, including the specific names of the subtypes and the distribution of patient cases within each, is presented in Table 4.1. An experienced pathologist meticulously curated masks for specific regions of WSIs that contained tumour tissues. The creation of these masks was based on careful examination of both the associated medical reports and the WSIs themselves. This process not only relied on the pathologist's expertise in identifying tumour regions but also on the valuable contextual information present in the diagnostic reports. By cross-referencing these two data sources – the WSIs, and the reports – a comprehensive set of tumour-centric masks was generated.

All WSIs in the dataset have been sectioned into patches of size 448 by 448 pixels. To ensure a balanced representation and to avert any potential biases, approximately 500 patches were extracted at the magnification of x20 from each WSI. This dense extraction process resulted in a total of roughly 86,500 patches derived solely from tumour regions. To further mitigate risks of data leakage and preserve the integrity of the dataset, the WSIs were partitioned into training, validation, and testing sets at the patient level. This measure ensures no patches from the same WSI appear across different sets. In the interest of robust and exhaustive experimentation, care was taken to include all 22 primary diagnoses in the testing phase, thereby ensuring the comprehensiveness of the conducted experiments.

PatchGastricADC22 [142]: This dataset contains 991 H&E-stained slides from distinct patients and has been procured from the surgical pathology files of the International

Primary Diagnosis	#WSI
Usual ductal hyperplasia	10
Columnar cell lesions including flat epithelial atypia	16
Atypical ductal hyperplasia	10
Sclerosing adenosis	11
Microglandular adenosis	3
Radial scar/complex sclerosing lesion	10
Tubular adenoma	6
Lactating adenoma	3
Intraductal papilloma	12
Encapsulated papillary carcinoma	4
Solid papillary carcinoma (in situ and invasive)	5
Atypical lobular hyperplasia	9
Lobular carcinoma in situ	10
Ductal carcinoma in situ	10
Metaplastic carcinoma	4
Invasive breast carcinoma of no special type	11
Invasive lobular carcinoma	11
Tubular carcinoma	4
Cribriform carcinoma	3
Mucinous carcinoma	11
Invasive micropapillary carcinoma	5
Carcinoma with apocrine differentiation	5

Table 4.1: Distribution of Primary Diagnosis: This table reports the number of each primary diagnosis in the GRH pathology dataset. This dataset encompasses 22 subtypes of breast cancer, establishing it as the most extensive and comprehensive breast cancer dataset compared with literature that predominantly considers two subtypes.

University of Health and Welfare, Mita Hospital, Tokyo, Japan. These slides were then digitized into WSIs at a magnification of 20x. All cases within this dataset have been diagnosed as adenocarcinoma and verified by three independent pathologists. The selection of these cases has been conducted randomly, reflecting a more accurate clinical distribution of adenocarcinoma subtypes. Accompanying diagnostic reports provided the source for histopathological captions.

These reports, initially in Japanese, have been expertly translated into English by two experienced pathologists. The translated descriptions composed a vocabulary of 277 words,

with the longest sentence not exceeding 50 words. As all cases have been sourced from the same hospital, text descriptions for cases with similar diagnoses displayed little variance, a trait mirrored in the translations.

In terms of image data, patches in size of 300 by 300 pixels have been primarily extracted from regions showcasing adenocarcinoma lesions, observed at a magnification of 20x. Initial loose annotations were carried out specifically on adenocarcinoma regions to minimize the extraction of non-adenocarcinoma regions. Owing to the variable sizes of the specimens and the adenocarcinoma regions within them, the number of patches extracted from each WSI varied. This variability also extended to the number of patches linked with a given caption. In the end, a total of 262,777 tiles (i.e., patches) have been extracted at a magnification of x20. In this study, the data is partitioned into training, validation, and test sets adhering to a consistent ratio: 70% of the data is allocated for training, 10% for validation, and the remaining 20% is reserved for the testing phase.

LC25000 [27]: A further dataset instrumental to the experiment is the LC2500. Comprising 25,000 patches, each with dimensions of 768 by 768 pixels, this dataset offers a substantial volume of high-resolution data for analysis, distributed evenly across five distinct classes. These classes each contain 5,000 images and are representative of the following histological entities: colon adenocarcinoma, benign colonic tissue, lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue.

This extensive dataset holds immense value for research due to several factors. Firstly, it adheres to strict standards of patient privacy, being fully de-identified and Health Insurance Portability and Accountability Act (HIPAA) compliant. Secondly, its quality is ensured through thorough validation processes. Furthermore, it is publicly accessible, encouraging open and collaborative research endeavours in the medical community.

The size and diversity of this dataset, combined with the high-resolution patches, in conjunction with its meticulous validation process, makes it particularly well-suited for rigorous experimental testing. In this study, the LC25000 dataset plays a crucial role in the experiment design and the results obtained therefrom.

4.4.2 Evaluation Metric

In order to comprehensively assess the efficacy of the model, as outlined in Section 3.8.2, the metric of **Recall at K** (R@k) is employed for both image-to-text and text-to-image retrieval tasks. However, given the intricacy and delicacy of the field of pathology, where more granular information can offer additional insight into the performance of the models, four different thresholds have been chosen for the value of K. Consequently, the model is

evaluated based on R@1, R@3, R@5 and R@10 across both retrieval directions, allowing for a multi-faceted appraisal of the model's capabilities.

Furthermore, to encapsulate the model's overall performance across these different thresholds, *R@sum* is computed and reported. This aggregate measure serves as a valuable tool in the comparison of various models, presenting a comprehensive picture of performance that effectively balances the results across different recall thresholds. This holistic perspective on model performance assists in drawing more nuanced conclusions to contribute to more informed decision-making.

4.4.3 Experiments

A comprehensive evaluation of the proposed method in comparison to alternative approaches was undertaken by applying two strategies tailored to the specific nature of the datasets. The datasets in question, namely GRH and LC25000, encompass images that have been classified according to primary diagnosis and can serve as a definitive classification criterion. This can help us to compare the proposed approach with other classification methods in the field, including KimiaNet [122]and HIPT [36] models. As a result, in the conducted experiments on these datasets, the primary diagnosis was utilized as a descriptor for the proposed method, and to facilitate a fair comparison with other techniques without compromising the task's generality, the primary diagnosis was mapped into labels compatible with other classification approaches.

Moreover, the proposed method was demonstrated to function as a *bi-directional retriever* for both GRH and LC25000 datasets by feeding the primary diagnosis as a description into the model. This allowed an evaluation of its performance in different settings, offering a more robust and comprehensive assessment of its efficiency and applicability.

In relation to the PatchGastricADC22 dataset, which is characterized by paired image-description entries at the WSI level, the provided descriptions were employed as text inputs. This was informed by the unique structure and specific attributes of this dataset, which is amenable to a deeper level of analysis courtesy of the substantial data encapsulated within the descriptions. Consequently, other existing cross-modal retrieval methodologies were used as points of reference for comparison and evaluation. This approach ensured that the evaluation benchmark remained constant, facilitating a rigorous comparison rooted in the specific characteristics of the dataset.

However, the dataset's structure limits in-depth result analysis. Evaluating the alignment of retrieved data points becomes intricate, especially when data from varying subtypes or classes align similarly. Additionally, pinpointing whether the model's identified

patterns align with pathologists' expectations is complex. Such in-depth evaluations would necessitate pathologist input.

In the subsequent sections, a thorough delineation of the procedural specifics associated with the methodology's implementation is provided, in addition to an in-depth presentation of the experimental results derived from each of the datasets. The aim of these discussions is to illuminate the complexities of the proposed approach, underscore its practical utility, and confirm its resilience when applied to diverse datasets and configurations.

GRH Benchmark Dataset: To conduct experiments on the GRH dataset, rigorous pre-processing steps were employed, where patches of size 448 by 448 pixels were extracted from WSIs based on Figure 4.3. The student model was trained using six stringent data augmentation techniques (τ_H), whereas the teacher model had a relatively relaxed regiment with two less intensive augmentations (τ_M) from the same patch as described in Section 4.2.2. Regardless of the specific augmentation technique implemented, the final processed image was adjusted to fit a uniform size of 224 by 224 pixels. Padding was added when necessary to ensure a consistent dimension across all images. The center crop of each path has been selected as the input to the LILE model. A similar augmentation set as τ_M is applied on this input for uniformity. The suite of augmentations applied to this dataset encompasses a range of transformations. These include Random Crop, Random Vertical and Horizontal Flip, as well as Color Jitter, which further encompasses changes in Brightness, Contrast, Saturation, and Hue. The intensity of these augmentations is strategically regulated, being categorized under either Hyper or Mild augmentation sets based on the severity of the transformation required. For the image representation extraction and SSL component, the student and teacher models utilized a ViT architecture similar to what was explained in Section 4.2.1 with 12 Transformer encoding blocks and 768 for the hidden states. Weight initialization for both student and teacher networks was facilitated through a pre-trained ViT model using the dataset featured in the "KimiaNet" paper [122]. The last six blocks of the student and teacher models are set to be trainable and frozen the rest.

The text encoder component is comprised of the versatile BioMed-RoBERTa-based network [66], with fine-tuning restricted to the last two blocks. An empirical approach led to the determination of optimal values for β and γ in Eq. 4.7, where the most desirable performance was observed at $\beta = 0.3$ and $\gamma = 1$. This configuration subtly de-emphasized the SSL component within the loss function. Given that the primary objective of the function is to identify and match paired images and texts, this adjustment allowed the model to focus more on its primary task while still benefiting from the auxiliary guidance of the SSL component.

Table 4.2: A comparison between the proposed method and other state-of-the-art approaches, examining their performance in both patch-based and WSI-based retrieval tasks on the GRH dataset.

Method	Text Retrieval				Image Retrieval				
	R@1	R@3	R@5	R@10	R@1	R@3	R@5	R@10	R@sum
Patch based									
KimiaNet [122]	24.1 ∓ 3.6	43.3 ∓ 2.3	51.2 ∓ 2.2	72.3 ∓ 1.9	N/A	N/A	N/A	N/A	N/A
BioMedCLIP [165]	12.6 ∓ 1.4	31.1 ∓ 3.7	42.7 ∓ 3.9	66.7 ∓ 5.3	18.4 ∓ 1.7	34.0 ∓ 3.2	41.9 ∓ 3.6	60.2 ∓ 4.2	307.6 ∓ 10.3
LILE [106]	29.3 ∓ 4.3	50.0 ∓ 3.5	63.6 ∓ 2.0	80.2 ∓ 1.9	39.2 ∓ 4.7	53.6 ∓ 4.2	57.5 ∓ 4.6	61.5 ∓ 3.1	434.9 ∓ 15.1
LILE+DINO	31.6 ∓ 4.2	53.8 ∓ 3.1	65.4 ∓ 2.5	81.9 ∓ 1.0	40.1 ∓ 6.7	57.6 ∓ 5.6	63.9 ∓ 4.8	64.8 ∓ 3.1	459.1 ∓ 23.7
LILE+H-DINO	$\mathbf{31.9 \mp 3.7}$	$\mathbf{54.2 \mp 3.1}$	$\mathbf{65.8 \mp 2.4}$	$\mathbf{82.3 \mp 1.0}$	$\mathbf{43.2 \mp 6.8}$	$\mathbf{59.8 \mp 4.1}$	$\mathbf{64.4 \mp 4.8}$	$\mathbf{65.9 \mp 3.5}$	$\mathbf{467.5 \mp 18.8}$
WSI based									
KimiaNet [122]	42.3 ∓ 5.3	66.3 ∓ 3.2	70.4 ∓ 3.0	78.1 ∓ 2.1	N/A	N/A	N/A	N/A	N/A
BioMedCLIP [165]	32.4 ∓ 2.1	48.4 ∓ 3.2	68.3 ∓ 3.4	74.1 ∓ 3.0	N/A	N/A	N/A	N/A	N/A
HIPT [36]	51.2 ∓ 6.8	72.4 ∓ 3.7	79.2 ∓ 2.1	86.5 ∓ 2.3	N/A	N/A	N/A	N/A	N/A
LILE [106]	50.0 ∓ 6.1	70.1 ∓ 4.2	76.4 ∓ 2.4	81.8 ∓ 2.9	N/A	N/A	N/A	N/A	N/A
LILE+DINO	53.7 ∓ 7.3	75.7 ∓ 4.3	83.2 ∓ 3.1	90.9 ∓ 2.6	N/A	N/A	N/A	N/A	N/A
LILE+H-DINO	$\mathbf{54.5 \mp 6.4}$	$\mathbf{77.3 \mp 4.5}$	$\mathbf{84.1 \mp 2.3}$	$\mathbf{92.5 \mp 3.3}$	N/A	N/A	N/A	N/A	N/A

To mitigate any potential bias and maintain the fairness of the comparison, six experimental iterations were carried out, each with a different test dataset and assigned the rest of the dataset as train and validation sets. To further validate these comparisons, it was ensured that each test set included all 22 primary diagnoses. Model training was accomplished using the AdamW optimizer [98], with an initial learning rate of $1e-6$. The learning rate was intelligently adapted to decay when the evaluation metric (R@sum) ceased to improve. The configuration for the rest of the model remains consistent with the optimal parameters identified during experiments with the ARCH dataset back in Chapter 3. This includes the number of iteration steps (denoted as K) and the weight factor (represented by α) within the loss function. The selection of these parameters is driven by the previous findings, thus ensuring that the model is effectively tuned for peak performance. This configuration continuity allows for a more robust evaluation and comparison across datasets. A comprehensive account of the findings from this setup can be found in Table 4.2.

The experiments on the GRH dataset were conducted in two different schemes. As it is shown in Table 4.2, the proposed method is compared with other methods in **patch-based** and **WSI-based** configurations. For the patch-based strategy, each patch needs to predict the correct primary diagnosis related to itself. In an alternate setup, the problem was approached through a WSI retrieval lens, replacing the individual patch-based retrieval. The strategy being used places considerable emphasis on **majority voting** for the prediction

of the WSI primary diagnosis. This approach, which is largely derived from strategies employed in patch-based experiments, is centred around the accumulation of predictions made for individual patches. The process unfolds as follows: for each patch, a determination is initially made regarding whether it has accurately retrieved its correct primary diagnosis, with various thresholds applied in the recall metric. If the true primary diagnosis is found among the top "K" retrieved diagnoses, the patch is identified as correctly retrieving the actual text. If not, the patch is assigned the label corresponding to the first retrieved primary diagnosis. Subsequently, a label is assigned to a WSI, reflecting the most common label among its constituent patches. This strategy provided a more holistic interpretation of the WSI, as it accounted for the collective intelligence of all patches in a WSI, rather than treating each patch in isolation. This perspective shift aimed at enhancing the overall retrieval performance by considering the consensus of predictions within a WSI, thus adding another dimension of robustness to the proposed method.

In the analysis of KimiaNet, a pre-trained model that had been subjected to comprehensive training using the vast TCGA dataset [4] was employed. This model underwent further fine-tuning using the training data, with the parameters of the last fully connected layer set to be trainable. The HIPT model was trained according to the guidelines provided in the authors' GitHub repository, designed to classify WSIs. For the experiments conducted on BioMedCLIP [165], their pre-trained model was utilized without any additional fine-tuning. Their model was trained on 15 million pairs of figures and text extracted from articles published on PubMed.

The proposed method, incorporating LILE structure into its backbone and employing the H-DINO strategy for self-supervision, demonstrated superior performance in both patch-based and WSI-based tasks compared to other approaches. A comparative analysis, pitching the proposed LILE+DINO method against the LILE without any self-supervision, demonstrated the potency of SSL in boosting model performance. The inclusion of SSL significantly elevated the R@sum metric, which measures the model's overall effectiveness, by a notable 32.6 margin. Furthermore, a comparison drawn between the LILE+H-DINO and the LILE+DINO methods underscored the value of implementing a patching method specifically devised for WSIs. This tailored approach to patching further enhanced the R@sum performance metric by an additional 8.4, highlighting its crucial role in optimizing the model's overall performance. This specialized approach notably improved the model's performance by providing a more robust embedding representation for the patches.

When considering text-to-WSI retrieval, the nature of the training and testing data, which are primarily patches, made it unfeasible to retrieve a WSI based on its primary diagnosis. This is due to the inherent fact that all patches extracted from the same WSI

share an identical primary diagnosis, leading to a uniform result. Consequently, no results have been reported in this direction, and this aspect is noted as "N/A" (Not Applicable). Further, it is worth mentioning that the KimiaNet and HIPT methods lack the capability to retrieve images when given text, leading to an absence of reported results for these models. However, within the scope of patch-based retrieval, the models that are built upon LILE display the compelling feature of bidirectional functionality. This characteristic highlights the distinct superiority of cross-modal retrieval models over traditional classification networks, such as HIPT and KimiaNet, as they possess the capacity to operate fluidly in both retrieval directions.

It should be noted that WSIs are predominantly utilized for diagnostic purposes by medical practitioners, including oncologists and pathologists. Consequently, greater emphasis is placed on the performance of the model in the WSI-based approach as compared to patch-based results. As seen in the table, the results for the WSI-based approach considerably outperformed those for the patch-based method. This phenomenon could be attributed to the power of majority voting. A WSI was assigned to a label if the majority of its patches voted for that specific label, implying that even if some patches were incorrectly predicted, the overall true label could still be accurately identified.

To further corroborate the efficacy of the proposed method, confusion matrices for R@1, R@3, and R@5 in patch-based strategy are presented in Figures 4.4, 4.5, and 4.6, respectively. Moreover, the comprehensive performance of the model in the context of WSI retrieval at various recall thresholds – R@1, R@3, and R@5 – is visualized in confusion matrices, as depicted in Figures 4.7, 4.8, and 4.9, respectively. To generate the confusion matrices that are displayed, the same methodology described in relation to Table 4.2 is employed. This approach allows for a consistent and comparative representation of the data across different visualizations.

These figures enable an in-depth understanding of the model's capability to assign diagnoses at different levels of recall accurately. From the confusion matrices can be understood that the proposed approach can distinguish more of its classes, and by going from R@1 to R@3 and R@5, i.e., looking at more retrievals, one can achieve more reliable predictions. Classifying 22 primary diagnoses of breast cancer is extremely challenging as they are all related to one anatomical organ, and they exhibit many similarities in their texture. As a result, R@3 and R@5 can provide more accurate results and be practical for pathologists and doctors to be relied on. Furthermore, it is noteworthy that each primary diagnosis can occasionally be misconstrued as a different primary diagnosis. This phenomenon could potentially serve as an additional source of information for pathologists. By allowing for the possibility of alternative diagnoses, insights that might be obscured if only the correct primary diagnosis were provided can instead be unveiled, thus supporting a more

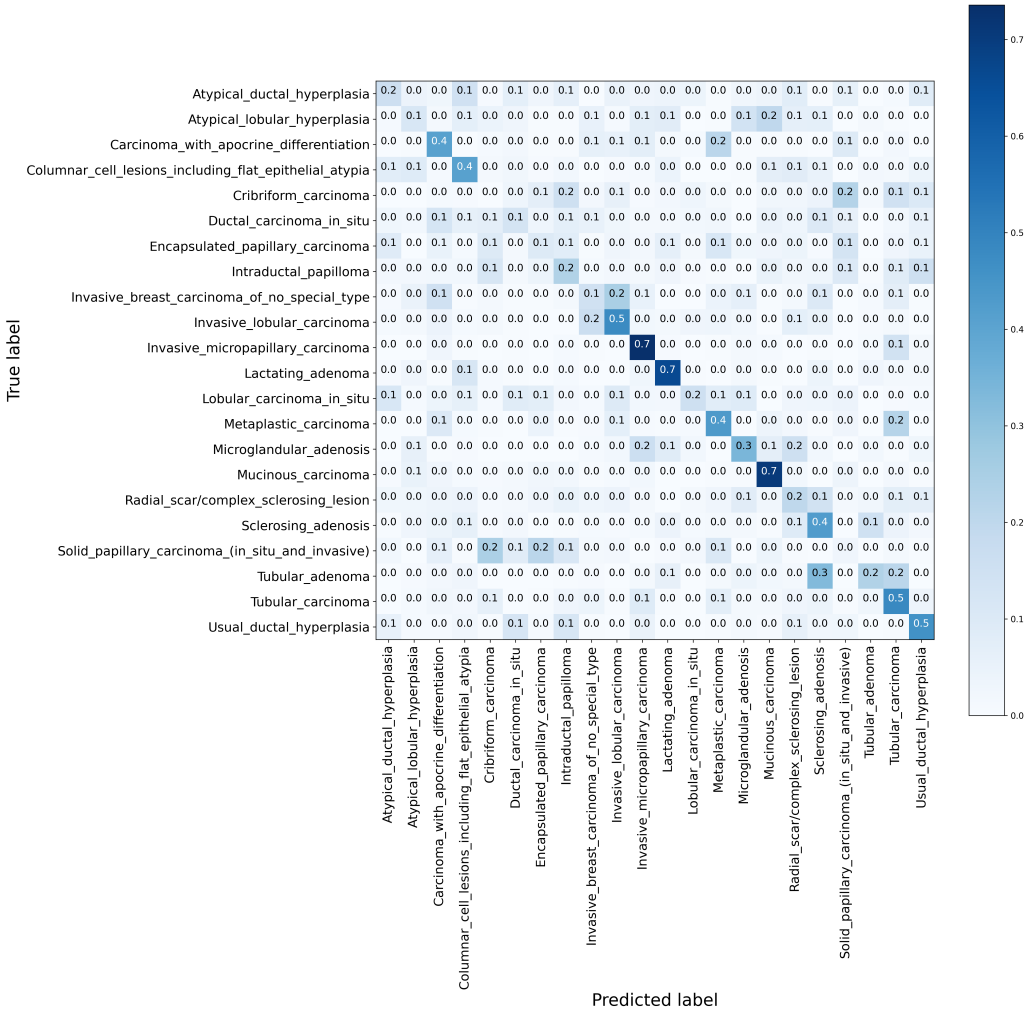


Figure 4.4: This confusion matrix depicts the retrieval performance of the proposed method for **patch-level** classification across 22 distinct primary diagnoses, specifically at an **R@1** recall level. It provides a comprehensive visual representation of how accurately the model classifies each diagnosis and demonstrates how the results are distributed across true and predicted classifications, hence offering a deeper understanding of the model's performance.

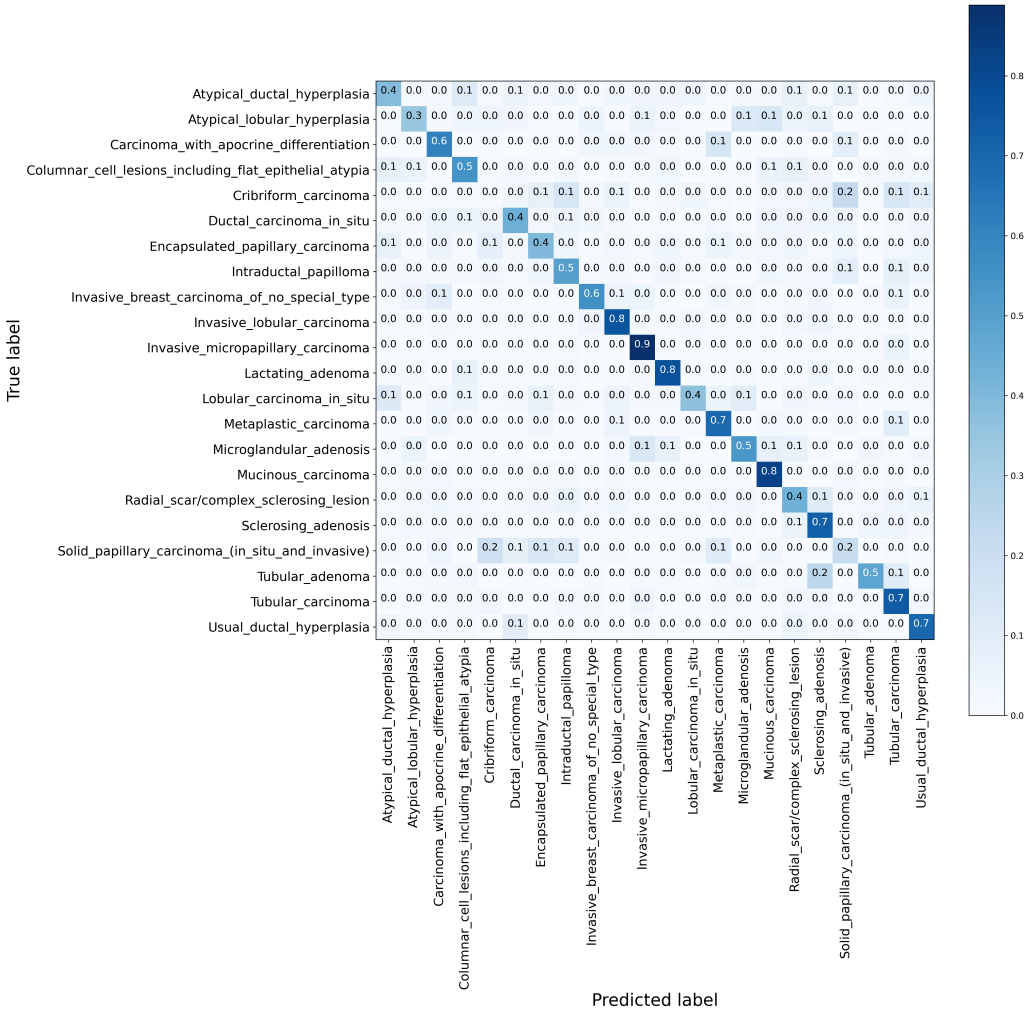


Figure 4.5: This confusion matrix depicts the retrieval performance of the proposed method for **patch-level** classification across 22 distinct primary diagnoses, specifically at an **R@3** recall level. It provides a comprehensive visual representation of how accurately the model classifies each diagnosis and demonstrates how the results are distributed across true and predicted classifications, hence offering a deeper understanding of the model's performance.

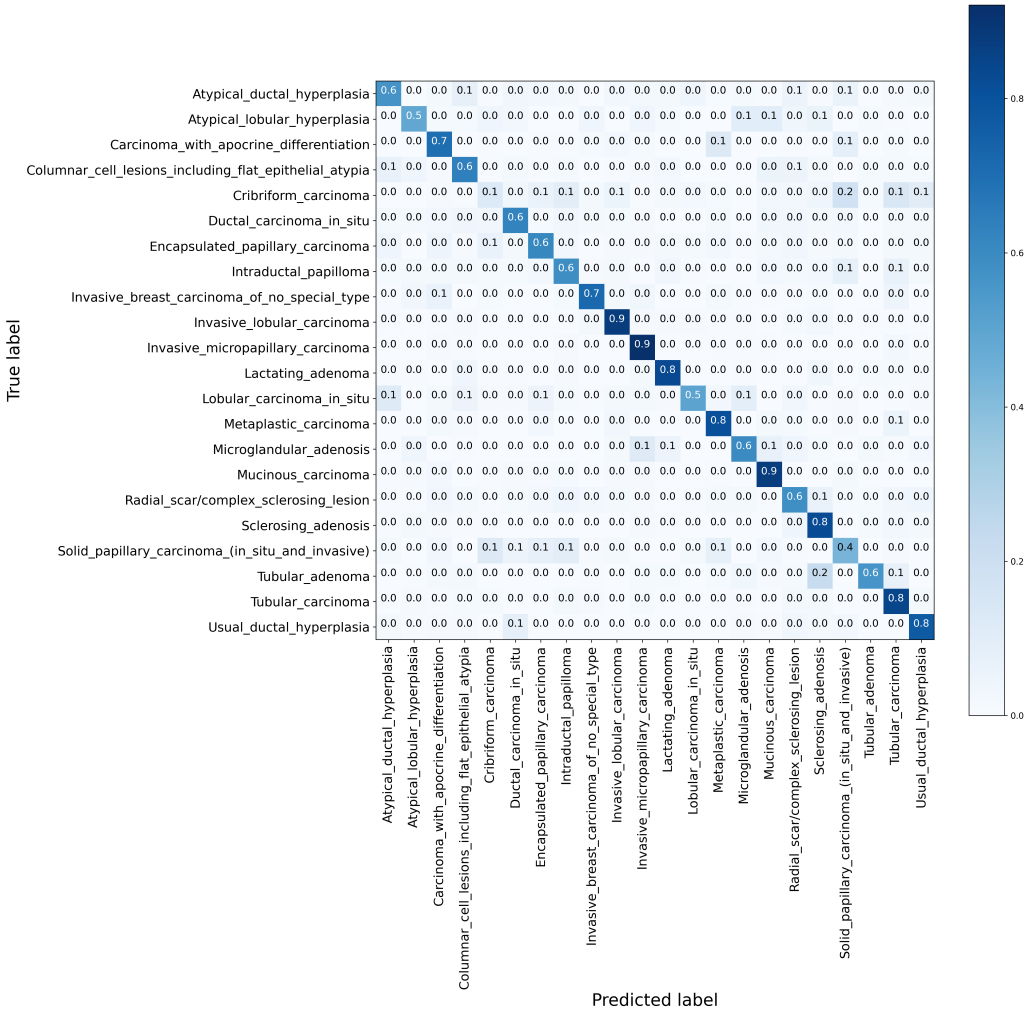


Figure 4.6: This confusion matrix depicts the retrieval performance of the proposed method for **patch-level** classification across 22 distinct primary diagnoses, specifically at an **R@5** recall level. It provides a comprehensive visual representation of how accurately the model classifies each diagnosis and demonstrates how the results are distributed across true and predicted classifications, hence offering a deeper understanding of the model's performance.

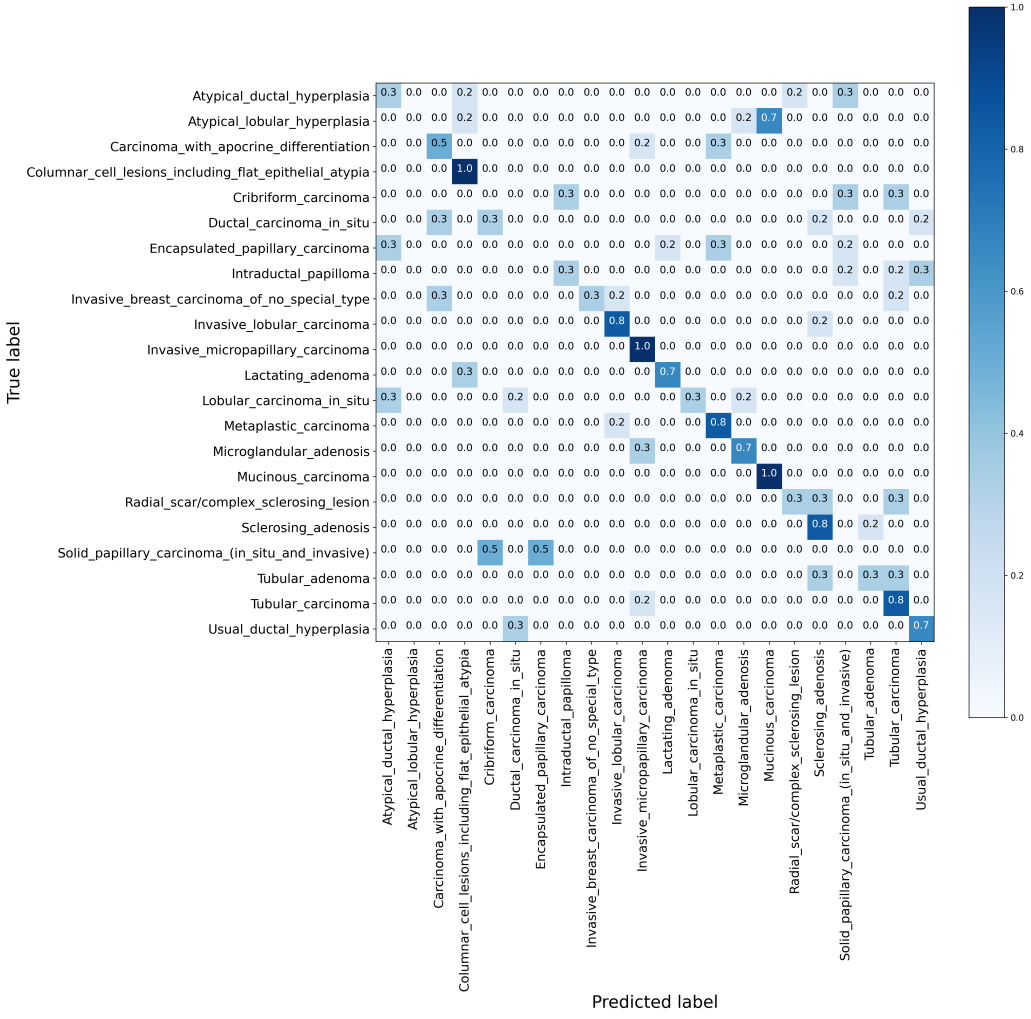


Figure 4.7: This confusion matrix provides a visualization of the retrieval performance at the **R@1** level of the proposed method for the **WSI-level** classification across 22 distinct primary diagnoses. It serves as a detailed display of the model's accuracy in diagnosis classification. By presenting the distribution of results across true and predicted classes, it enables a profound understanding of the overall efficacy and precision of the model.

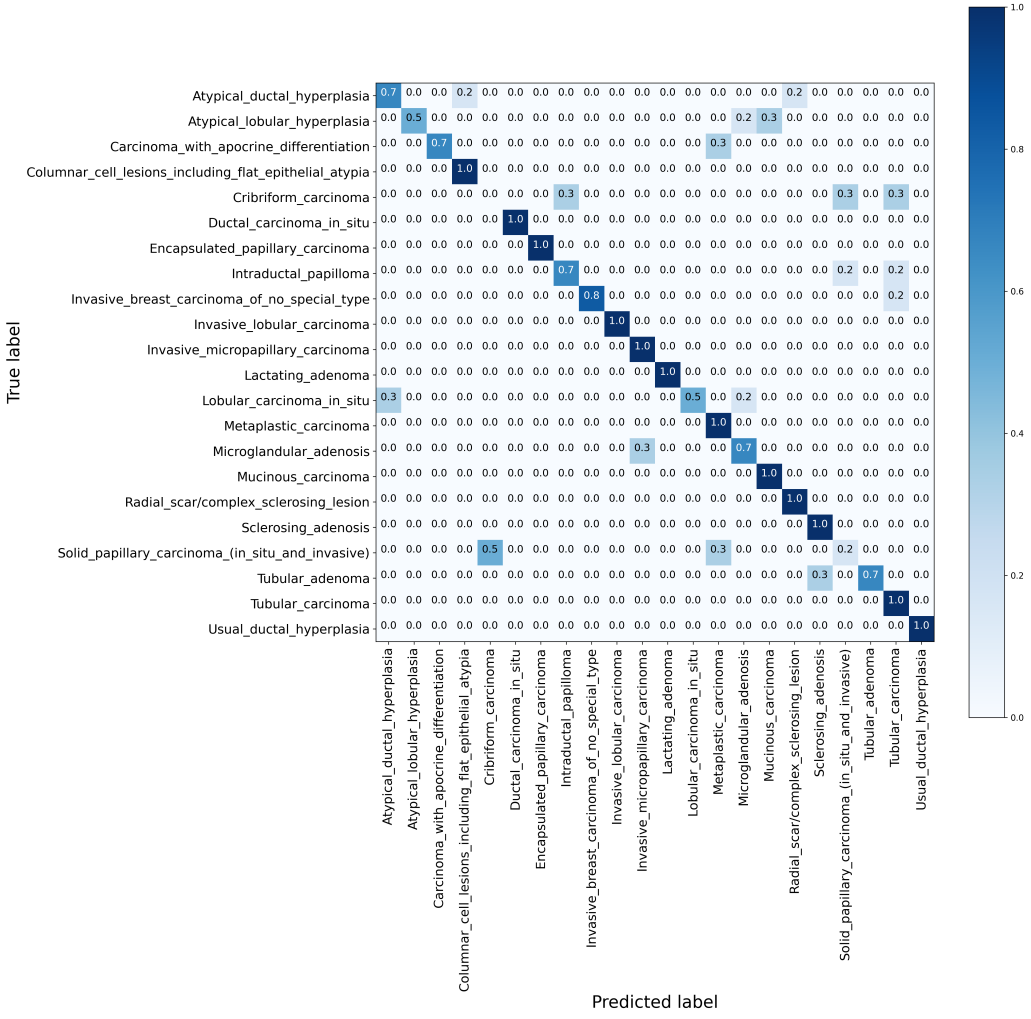


Figure 4.8: This confusion matrix provides a visualization of the retrieval performance at the **R@3** level of the proposed method for the **WSI-level** classification across 22 distinct primary diagnoses. It serves as a detailed display of the model's accuracy in diagnosis classification. By presenting the distribution of results across true and predicted classes, it enables a profound understanding of the overall efficacy and precision of the model.

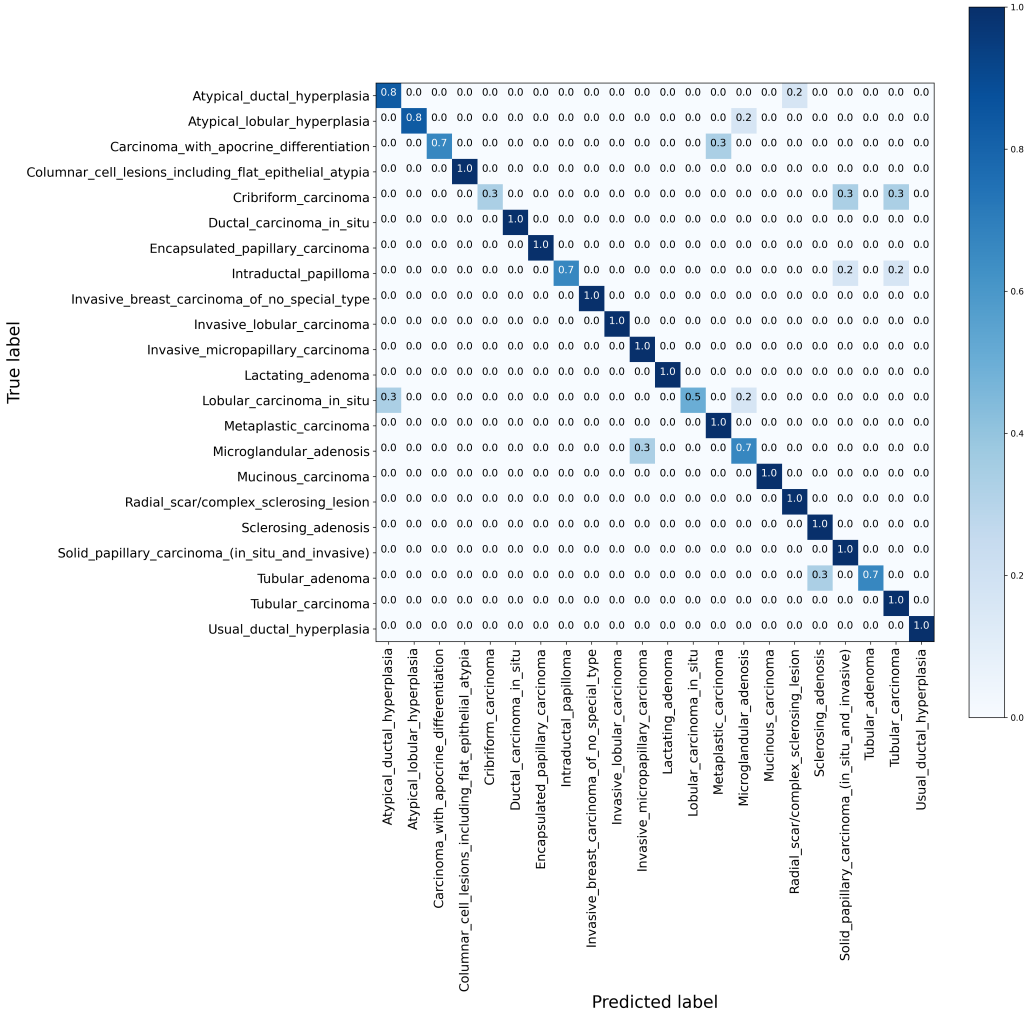


Figure 4.9: This confusion matrix provides a visualization of the retrieval performance at the **R@5** level of the proposed method for the **WSI-level** classification across 22 distinct primary diagnoses. It serves as a detailed display of the model's accuracy in diagnosis classification. By presenting the distribution of results across true and predicted classes, it enables a profound understanding of the overall efficacy and precision of the model.

comprehensive understanding of the pathology in question.

More examination of the confusion matrices reveals that two primary diagnoses, namely, *Caribiform Carcinoma* and *Lobular Carcinoma in Situ*, pose significant challenges. A contributing factor to the misclassification of Caribiform Carcinoma pertains to the limited representation of this diagnosis in the GRH dataset. With only three WSIs corresponding to this diagnosis, the dataset configuration for each training run is such that one WSI is set aside for testing, and another for validation, leaving just a single WSI for training. As depicted in Figure 4.10, which showcases three random patches extracted from these three WSIs, the structural and feature variation across each WSI is quite significant. Consequently, training the model using a single WSI fails to provide a comprehensive representation of the primary diagnosis, thereby making it challenging to correctly identify this diagnosis in other WSIs.

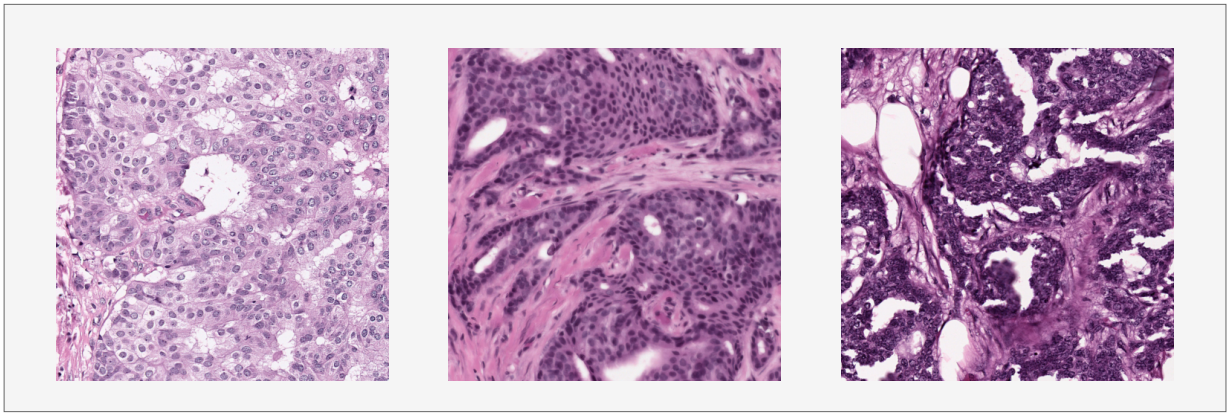


Figure 4.10: Three randomly selected patches from three distinct Cribriform Carcinoma WSIs present in the GRH dataset. The noticeable differences in structure and the significant variations among these images illuminate the challenge of accurate classification, ultimately contributing to lower accuracy rates.

If more data corresponding to this specific primary diagnosis were to be included in the training set, it would be reasonable to anticipate improved diagnostic accuracy. By enhancing the diversity and volume of the training data, the model would be equipped with a more holistic understanding of the diagnosis, thus improving its ability to identify and classify new instances accurately.

In another experiment, the robustness and adaptability of the proposed model were demonstrated through an analysis of the generated embeddings and the real-world application of these findings. The dataset used for this experiment was sourced from the WHO

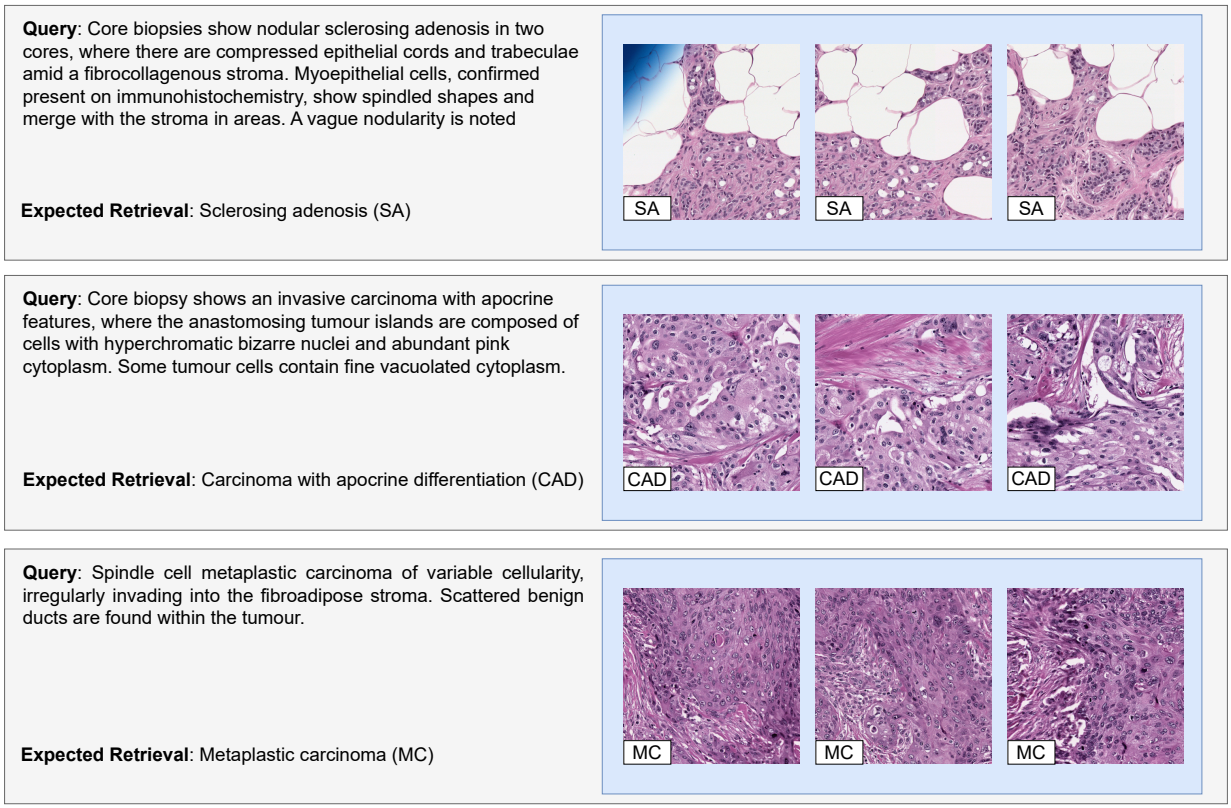


Figure 4.11: Visualizing image retrieval using real-world medical descriptions extracted from the WHO dataset, which wasn't included in the original training set. Under each description, the associated primary diagnosis is displayed. To the right, the top three images retrieved by the model are presented alongside their corresponding primary diagnoses. This illustration showcases the model's ability to retrieve and match images based on textual medical descriptions accurately.

(World Health Organization), which was not part of the original training set but provided real medical reports all related to the same primary diagnosis.

The top three results from the text-to-image (t2i) retrieval process are displayed in Figure 4.11. The capability of the model to retrieve accurate and relevant images given a description of a specific primary diagnosis is clearly illustrated by these results.

These examples demonstrate how the model can identify and understand similar semantic structures between images and their corresponding descriptions. Not only does this enhance the overall precision of the model, but it also demonstrates its ability to adapt to

Table 4.3: Comparative analysis of methodological performance on the GRH dataset using Recall@K@V metric with a "V" threshold of five - illustrating the superior performance and adaptability of the proposed LILE+H-DINO method across diverse recall thresholds.

Method	WSI based			
	R@1	R@3	R@5	R@10
KimiaNet [122]	43.0 ∓ 5.0	63.6 ∓ 2.3	69.1 ∓ 2.5	78.5 ∓ 2.3
HIPT[1] [36]	51.2 ∓ 6.8	72.4 ∓ 3.7	79.2 ∓ 2.1	86.5 ∓ 2.3
LILE [106]	50.8 ∓ 6.2	70.0 ∓ 5.6	76.0 ∓ 1.8	81.5 ∓ 2.4
LILE+DINO	56.6 ∓ 5.5	72.5 ∓ 5.4	80.9 ∓ 3.2	89.9 ∓ 3.5
LILE+H-DINO	$\mathbf{57.2 \mp 5.2}$	$\mathbf{74.1} \mp 4.7$	$\mathbf{81.8} \mp 4.0$	$\mathbf{91.9} \mp 4.0$

and comprehend new, intricate concepts.

In additional sets of experiments conducted on the GRH dataset, a focus was placed on evaluating the model based on metrics more aligned with the diagnostic considerations of pathologists in order to ensure real-world clinical applications were closely reflected.

One of the key metrics, recall@K (R@k), which was reported previously, was utilized for this evaluation. The successful retrieval was defined by the premise that the correct item was found within the top K retrieved items. This measure of success was then used to cast a 'vote' among the patches of a WSI.

Taking the approach a step further, another strategy was adopted. This method involved the top "V" retrieved primary diagnoses for each image patch being gathered. Here, "V" functions as a voting threshold, representing the number of retrieved items deemed informative enough to contribute to the majority vote in the ultimate step. Subsequently, these items were multiplied by their corresponding normalized similarity scores. The primary diagnoses retrieved for each WSI were then organized based on the frequency of their occurrence among the top "V" diagnoses, which were weighted by their scores across all patches within that specific WSI. Following this, the recall@K evaluation was applied to ascertain if the accurate primary diagnosis was among the top "K" retrieved items. This layered evaluation technique, termed **Recall@K@V**, provided a multidimensional and more nuanced analysis of retrieval success and presented a more comprehensive picture of retrieval performance. Such a metric can be of assistance to pathologists when ordering immunohistochemistry; Recall@K@V is a good quantification to select the right biomarkers.

[1]In this experiment, given that the HIPT method isn't adaptable to the proposed voting approach, the results for WSI classification have been reported.

This innovative and advanced evaluation strategy facilitated the procurement of a more detailed and nuanced understanding of the model's performance. The results of this experiment, which offer an extensive analysis of Recall@K@V evaluations, are depicted in Figure 4.12. The strategic expansion of the traditional approach allowed for a thorough and informative dissection of the model's performance.

Figure 4.12 illustrates the impact of modulating the voting threshold, termed "V", on the task of determining a WSI's primary diagnosis over a range of Recall@K values. The voting threshold specifies the number of top-retrieved items per patch taken into account in the decision-making process. By adjusting "V", the breadth of information contributing to the final diagnostic decision can be fine-tuned, offering a nuanced perspective on the retrieval performance. The visual representation shows an initial trend of improved performance with an increasing voting threshold, underlining the value of a more inclusive decision-making process. Multiple "V" thresholds were examined to investigate this relationship, ranging from a conservative limit of 1 (i.e., only the top-retrieved item per patch contributes to the final decision) to a comprehensive limit of 22, encompassing all primary diagnoses in the database. The effects of these different thresholds are colour-coded for ease of comparison.

However, it is critical to note that there appears to be a limit to the benefit of this expansion. As demonstrated by the figure, a voting threshold greater than five signifies the beginning of an observable performance decline. This observed decline aligns with expectations, considering that the similarity scores across primary diagnoses tend to cluster after the top few candidates. Essentially, including too many lower-ranked diagnoses dilutes the influence of the most probable ones, leading to less accurate final decisions. Thus, a careful balance is required when choosing the voting threshold to optimally leverage the multi-voting scheme's advantages while avoiding the pitfalls of over-inclusion.

Continuing to add votes past the optimal threshold tends to introduce excess noise into the voting system. This "over-voting" counter-productively decreases the precision of the voting mechanism and consequently reduces the final recall score. Hence, while expanding the voting pool can initially enhance diagnostic accuracy, careful moderation is necessary to prevent a decline in performance. The comparative performance of different methodologies on the GRH dataset, as measured by this particular metric, is detailed in Table 4.3. The "V" threshold within the Recall@K@V metric has been set to five for these evaluations. This setting was chosen as it consistently delivered the most advantageous results across different trials, as seen in Figure 4.12. A clear observation from these comparisons is that the proposed LILE+H-DINO method excels beyond other approaches across varying recall thresholds. Notably, its superior performance holds up regardless of the recall threshold applied, further underscoring the robustness and generalization capability of the model.

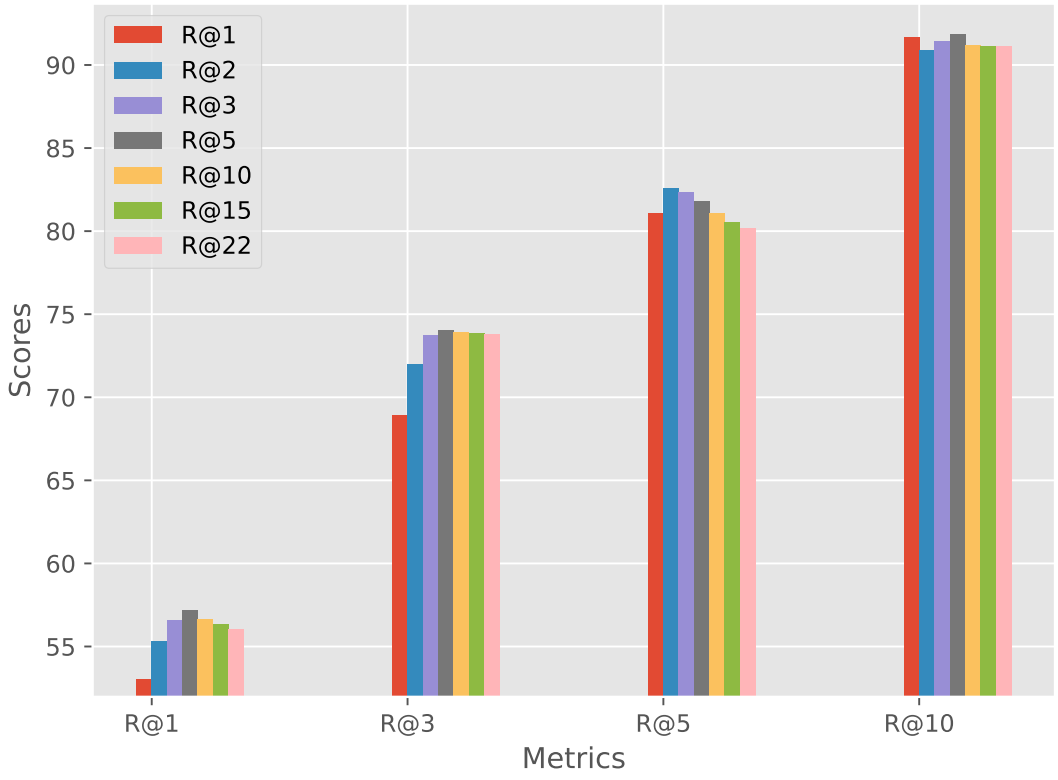


Figure 4.12: This figure illustrates the Recall@K@V values for various combinations of V and K. In this process, the most relevant primary diagnoses are first sorted for each patch based on a given V value. Subsequently, the Recall@K is computed for each WSI using the frequency of the most common primary diagnosis identified for that WSI. The results of this process are showcased in the figure for multiple voting thresholds (V), which include 1, 2, 3, 5, 10, 15, and 22. The selection of these thresholds is significant as the number of primary diagnoses is capped at 22. Additionally, the figure also displays these results across varying levels of Recall (K), specifically 1, 3, 5, and 10. This visualization provides a comprehensive overview of how alterations in the voting threshold and recall level influence the Recall@K@V values, thereby highlighting the nuanced interplay of these parameters in the evaluation process.

This reliable higher performance of the LILE+H-DINO model illustrates its potency and adaptability across diverse retrieval contexts in the GRH dataset.

PatchGastricADC22 Benchmark Dataset: In the processing of the PatchGastri-

cADC22 dataset, patches of 300 by 300 pixels in size were utilized. The description of each extracted patch was linked to the corresponding WSI it was obtained from. The primary distinction in this approach is manifested in the input selection for the model, which acts as an anchor for the training of the retrieval part. Instead of a center crop being traditionally employed, variability is introduced: for each iteration in the training process, a random crop of dimensions 224 by 224 pixels is chosen from the original 300 by 300 patch and inputted into the model. This methodology is based on the assumption that a subset of the original patch, specifically the 224 by 224 pixels crop, can adequately represent the accompanying descriptive text. Moreover, this method infuses a layer of randomness into the training process, potentially fostering an enhanced generalization ability for the model. However, during the test phase, a center crop is utilized. The overall model architecture was maintained in alignment with the configuration adopted for the GRH experiments. Pre-trained models from a classification task using the dataset featured in the "KimiaNet" paper [122] were applied to both the student and teacher networks, while the BioMed-RoBERTa was employed as the text encoder [66]. The only variable in this arrangement was the number of trainable blocks in the vision component, where the last eight blocks were selected for adjustment. For the text encoder, the last two blocks were set as trainable. A learning rate of $5e-5$ was chosen, and the batch size was set at 64.

Random cropping with a size of 224 by 224 for extracted different views is applied. The data augmentation methodology implemented with the GRH dataset was mirrored here, with six Hyper augmentations (τ_H) being applied to the student model and two less stringent ones (τ_M) to the teacher model. These augmentations include Random Crop, Random Vertical and Horizontal Flip, and Color Jitter (affecting Brightness, Contrast, Saturation, and Hue). The strength of these transformations was controlled, falling into either Hyper or Mild augmentation sets, depending on the desired extent of change. Optimal results were achieved when the parameters of the loss function were finely tuned, setting γ to 1 and β to 0.5. Given that the central goal of the model is to effectively locate and pair corresponding images and texts, this particular adjustment aids the model in concentrating on its main task.

Table 4.4 documents the performance of the model in patch-based and WSI-based retrieval compared to other methods. This table provides insights into the efficacy of the suggested method and its relative performance in comparison with other techniques.

To evaluate the performance of the proposed approach and compare it with other methods, LILE is used as the baseline and compared LILE+H-DINO with LILE+DINO and BioMedCLIP [165] which is one of the *foundation models* for vision and vision language tasks that trained on large paired medical data. Notably, there are no pre-existing reported results for this dataset, which underscores the novelty of this research. Furthermore, for

Table 4.4: A comparison of the proposed method and other state-of-the-art approaches on the PatchGastricADC22 dataset, evaluating their performance in patch-based and WSI-based retrieval tasks.

Method	Text Retrieval				Image Retrieval				
	R@1	R@3	R@5	R@10	R@1	R@3	R@5	R@10	R@sum
					Patch based				
BioMedCLIP (zero-shot) [165]	6.0	17.8	28.7	53.2	10.1	22.4	29.8	51.2	219.2
LILE [106]	18.2	37.8	49.1	78.2	28.5	46.9	48.3	68.2	375.2
LILE+DINO	19.6	40.4	53.8	80.0	34.2	50.6	55.0	70.0	403.5
LILE+H-DINO	**20.8**	**42.2**	**54.8**	**81.3**	**40.0**	**55.0**	**55.0**	**75.0**	**424.1**
					WSI based				
BioMedCLIP (zero-shot) [165]	14.6	26.2	32.3	64.5	N/A	N/A	N/A	N/A	N/A
LILE [106]	32.4	46.1	64.3	85.8	N/A	N/A	N/A	N/A	N/A
LILE+DINO	35.0	49.5	66.2	88.3	N/A	N/A	N/A	N/A	N/A
LILE+H-DINO	**36.7**	**52.3**	**67.8**	**93.2**	N/A	N/A	N/A	N/A	N/A

the sake of comparison, only the BioMedCLIP model was utilized in this instance due to its public availability and its standing as one of the SOTA techniques for cross-modal retrieval tasks in the medical field.

The same voting scheme proposed for the GRH dataset is applied for this dataset with a voting threshold of five for WSI-based retrieval tasks. The results show the efficacy of the combination of LILE with H-DINO in both patch-based and WSI-based tasks which outperformed all the other approaches. In this experiment, the pre-trained BioMedCLIP model [165], which was trained on a large volume of figure-caption pairs extracted from biomedical research articles in PubMed Central, was employed without further fine-tuning. To facilitate the application of the BioMedCLIP model for WSI-based retrieval, the identical procedure used for LILE and other methods was diligently implemented. The results of BioMedCLIP, as expected, are the lowest among other approaches as it is only trained with images and text extracted from research articles, and it can show the importance of data in training models. The comparative evaluation of the proposed method, LILE+H-DINO, with LILE and LILE+DINO, offers some notable insights.

Among the methods charted in the table, a distinct performance advantage was exhibited by the proposed LILE+H-DINO. It outpaced both LILE and LILE+DINO at various recall thresholds, including R@sum, registering a lead of 48.9 and 20.6, respectively. This performance margin underscores the significance of incorporating self-supervised learning

(SSL) and a tailored patching scheme in the proposed methodology, elements that played a pivotal role in enhancing retrieval accuracy.

On another note, a characteristic of this dataset is that each description is associated with all patches extracted from the corresponding Whole Slide Image (WSI). Consequently, retrieving a WSI from a description is not feasible, as the correlation doesn't lend itself to direct WSI retrieval. Therefore, the corresponding entries in Table 4.4 are marked as "N/A" (Not Applicable), indicating the inapplicability of certain metrics in this context.

The dataset under consideration contains authentic captions penned by pathologists, offering a practical application scenario for the proposed approach. Furthermore, the challenge escalates when distinguishing between descriptions and images pertaining to the same type of adenocarcinoma. The similarity in the nature of the images and text can blur the differentiation, thereby posing a considerable challenge. This is especially prominent when comparing the performance to that of the GRH dataset. The complex task of accurately matching the nuanced visual patterns in the images with the precise medical terminology in the captions can lead to reduced performance, reinforcing the intrinsic difficulty of the task. Nonetheless, successfully tackling this challenge can provide valuable insights for advancing cross-modal retrieval techniques in real-world clinical settings.

LC25000 Benchmark Dataset: The LC25000 dataset was employed as another benchmark to evaluate the effectiveness of the proposed methodology. Patches of dimensions 768 by 768 pixels characterize this particular dataset. In line with model requirements, these patches were resized to 448 by 448 pixels. The methodology previously detailed for the GRH dataset was adhered to, ensuring consistent operations across all datasets. As mentioned for the PatchGastricADC22 dataset, instead of the center 224 by 224 pixels crop feed into the model, a random crop has been fed to the cross-modal retrieval network as the same characteristic applied for the LC25000 dataset. One limitation that arose with the LC25000 dataset was the absence of information regarding the WSI from which the patches were extracted. This resulted in the reporting of *only patch-based retrieval outcomes* for this dataset.

The preparation of views for both the student and teacher networks was undertaken in an identical manner to previous operations for the GRH and PatchGastricADC22 datasets, including the number of views and augmentations that have been applied. An exhaustive and systematic experimental analysis has ascertained optimal parameter settings tailored for this dataset. In the context of the image encoder, the student and teacher networks, fine-tuned on the last six blocks of a pre-trained ViT trained on the dataset featured in the "KimiaNet" paper [122], showed the best results. On the other hand, the last two blocks

of BioMed-RoBERTa were found to be most effective for the text encoder.

In terms of optimizing the loss function, the parameters were carefully calibrated. The parameter γ was set to 1, while the parameter β was set to 0.4, granting a slightly lesser emphasis to SSL part in the loss function. A learning rate of $5e-5$ was employed, offering a good balance between training speed and model stability. The batch size was 64.

Comparisons were made between the performance of the proposed method and other existing methods. The insights derived from these comparisons have been collated and presented in Table 4.5. The findings affirm the proposed method's robustness and effectiveness when applied to this benchmark dataset, supporting its validity across various scenarios. Notably, the MD [107], CNN and SVM [148], and MRFO & EO [138] methods are unable to be applied in the image retrieval direction, and they do not provide results for R@3 and R@5. The absence of reported results for these models further highlights this limitation. The outcomes of this experiment are noteworthy and closely matched, largely due to the straightforward nature of the task, which involves recognition among merely five distinct categories. In this context, the distinguishing factor among various methodologies primarily lies in the R@1 metric (the strict classification case), given that all methods achieved a perfect score of 100 for R@3 and R@5. Consequently, R@10 has not been reported, considering that with only five primary diagnoses, any recall threshold above 5 would invariably yield a score of 100.

Despite the high level of competition, the proposed model, LILE+H-DINO, manages to inch ahead, albeit by a narrow margin. It surpasses the performance of MRFO & EO, as cited in [138], by 0.2 for R@1, and outperforms LILE and LILE+DINO by 0.5 and 0.1, respectively, for R@1, which subsequently impacts the R@sum score.

This marginal yet significant superiority is particularly noteworthy given the circumstance where scores are concentrated near the ideal score of 100. The model's ability to offer this subtle performance boost in a context where the scope for enhancement is minimal indicates its capacity to yield exceptional results in more complex scenarios.

Furthermore, image retrieval results have only been provided for LILE architectures. This is because other approaches have only been capable of functioning in a single direction - specifically, classifying images based on their primary diagnosis. This further highlights the versatility and capability of LILE networks in the context of image retrieval tasks.

Table 4.5: A comparison of the proposed method and other state-of-the-art approaches on the LC25000 dataset, evaluating their performance in patch-based retrieval tasks.

Method	Text Retrieval			Image Retrieval			
	R@1	R@3	R@5	R@1	R@3	R@5	R@sum
	Patch based						
MD [107]	98.4	N/A	N/A	N/A	N/A	N/A	N/A
CNN and SVM [148]	94.0	N/A	N/A	N/A	N/A	N/A	N/A
MRFO & EO [138]	99.6	N/A	N/A	N/A	N/A	N/A	N/A
LILE [106]	99.3	100	100	100	100	100	599.3
LILE+DINO	99.7	100	100	100	100	100	599.7
LILE+H-DINO	**99.8**	**100**	**100**	**100**	**100**	**100**	**599.8**

4.5 Summary

In this chapter, the innovative approach of LILE+H-DINO has been introduced. This extended model, which uniquely integrates the cross-modal retrieval architecture with SSL tailored for the vision encoder, is distinguished by a distinctive patching strategy known as *harmonizing DINO*, short H-DINO. Furthermore, the model is trained using an end-to-end scheme, allowing for seamless interactions between the components of the model.

The challenge of securing labelled images in the medical field has been acknowledged. In response to this, the application of SSL has been limited solely to the vision encoder, an approach that strategically addresses the issue. The unique combination of LILE and SSL has been demonstrated to improve image representation, which subsequently leads to enhanced performance in cross-modal retrieval tasks.

A rigorous evaluation of the model has been conducted across various medical benchmark datasets. The results achieved have quantified the superior capabilities of the proposed approach, showing that it consistently outperforms other models in a range of scenarios and across different performance metrics.

This innovative endeavour in utilizing SSL specifically for the vision encoder not only offers promising improvements in model performance but also underscores the potential of such hybrid methodologies in overcoming complex challenges within the medical image retrieval domain. The end-to-end training scheme of the model further ensures the effective interaction of different components and contributes to the overall success of the approach.

Additionally, the versatility of this methodology extends beyond its current application; it can be adapted to accommodate various data types or staining techniques, enabling the retrieval of diverse modalities or different staining patterns. To achieve this, one would need to compile a dataset that includes these varied modalities or staining techniques. Subsequently, the model would require re-training based on the new dataset to fine-tune its performance for these specific tasks.

Chapter 5

Summary and Future Work

5.1 Summary

The facilitation of efficient and versatile retrieval across varied modalities, such as images and texts, is the core goal of cross-modal retrieval tasks for many real-world applications. These tasks can be more challenging when the labelled data are scarce. To address this problem, this thesis proposed a method that can extract robust embeddings for both image and text modalities. The unique element of the suggested approach is the integration of SSL, applied within the scope of knowledge distillation to cross-modal retrieval tasks and the execution of an end-to-end training regime.

In the proposed paradigm, an effort was made to extract the most significant feature embeddings from the given images via SSL, with a particular emphasis on the recently emerged DINO architecture. This approach is further optimized by introducing a novel patching scheme named *harmonizing DINO*, short H-DINO, which is tailored for pathology WSIs. Consequently, a refined methodology is developed that progressively identifies and captures the most important features across modalities, both in isolation and relative to one another.

A multi-head self-attention module was employed, thereby enhancing the network's understanding of its assigned modality. Concurrently, the output from this module is directed into a cross-attention module, which assists the model in locating potential alignments between the two modalities, namely text and image. This crucial step ensures that the features retrieved from one modality can be fine-tuned by considering the influence of the other. An iterative alignment scheme grounded in a memory network is employed in

the proposed LILE+H-DINO strategy, serving to augment the retrieved features based on significant data segments highlighted by the self-attention and cross-attention modules.

Moreover, a novel loss function was introduced that encompasses the objectives of both the SSL and cross-modal retrieval tasks. Weights were utilized in this function to adjust the significance of each objective group. This strategic weighting helps achieve a balance between these distinct but interconnected objectives.

A rigorous assessment of the efficiency of the LILE+H-DINO methodology was conducted across various datasets, including MS-COCO, ARCH, GRH, PatchGastricADC22, and LC25000. Consistent superiority was demonstrated by the proposed approach in terms of the *Recall@K* metric, a standard measure for retrieval tasks in both general purpose and histopathology contexts. In addition, excellent performance was observed with the introduced *Recall@K@V* metric, specifically designed to capture the nuances of multi-class retrieval tasks in the histopathology field. Notably, the pioneering attempt to address the challenge of bidirectional cross-modality retrieval in digital pathology was marked by the proposed method, highlighting its potential to instigate transformative changes in this field.

The effectiveness of an SSL scheme applied to cross-modal retrieval tasks was clearly showcased by the proposed method. Similar images, although from different views, were forced to adopt similar embeddings, enabling more insightful representations to be extracted. This approach takes a step further by integrating a unique patching scheme explicitly devised with WSIs in mind. Harmonizing patching is a departure from the local-global concept in DINO literature to accommodate the anatomic intricacies of histopathology. This tailored patching strategy utilizes the expansion of the chosen patch while maintaining a consistent scale, i.e., the same magnification in digital pathology. Such a practice ensures that the contextual integrity and structural consistency of the patch are preserved, thus contributing to a richer and more accurate interpretation of the image content. The distinctions in performance between LILE, LILE+DINO, and LILE+H-DINO illustrate the advantages of merging SSL and the tailored patching scheme.

The importance and relevance of the proposed methodology are underscored by the diversity of the datasets utilized in the experiments. Encompassing 22 primary diagnoses, the GRH breast cancer dataset stands as the most formidable barrier encountered in this research, thus offering a significant challenge due to its diagnostic breadth, and almost imperceptible anatomic subtleties. The proposed method's practical application is further emphasized by its deployment on a dataset consisting of images and descriptions composed by pathologists. This implementation demonstrates the real-world applicability of the method, emphasizing its practical potential. On the other hand, the model's performance

was also tested using a less challenging dataset that included only five primary diagnoses. This test confirms the method's adaptability to various scenarios, ensuring its robustness and broad applicability.

5.2 Future Works

The comprehensive study presented in this thesis provides multiple avenues for further exploration and enhancement in the field of cross-modal retrieval tasks.

1. **Extend to Other Modalities**: While the presented work has largely focused on image-text cross-modal retrieval, the methodology can be extended to include other modalities such as genome data and even other image modalities including Magnetic resonance imaging (MRI) or Computed tomography (CT) images form the radiology domain. This expansion can pave the way for more comprehensive and versatile multimodal systems.

2. **DINO for Other Applications**: This study demonstrated the effective use of DINO for image embeddings. Further research could investigate how DINO and similar SSL methods could be applied to other tasks within the realm of medical image analysis, such as segmentation or classification.

3. **Explainability and Interpretability**: As AI models become more sophisticated and are increasingly used in critical applications like healthcare, the need for explainability and interpretability becomes crucial. Future research could focus on developing techniques that make the cross-modal retrieval process more transparent and interpretable.

4. **Large-scale and Real-world Applications**: The methodologies proposed in this thesis should be validated and further improved upon by using larger and more diverse datasets. In addition, this research could focus on the practical implementation of these techniques in real-world medical applications, studying their impacts on clinical decision-making, workflow, and ultimately patient outcomes.

References

[1] English wikipedia - wikipedia. https://en.wikipedia.org/wiki/English_Wikipedia. (Accessed on 11/04/2021).

[2] The fall of rnn / lstm. we fell for recurrent neural networks... — by eugenio culurciello — towards data science. https://towardsdatascience.com/the-fall-of-rnn-lstm-2d1594c74ce0. (Accessed on 11/11/2021).

[3] Gdc. https://portal.gdc.cancer.gov/. (Accessed on 11/13/2021).

[4] Gdc. https://portal.gdc.cancer.gov/. (Accessed on 07/23/2023).

[5] Github - jcpeterson/openwebtext: Open clone of openai's unreleased webtext dataset scraper. this version uses pushshift.io files instead of the api for speed. https://github.com/jcpeterson/openwebtext. (Accessed on 11/04/2021).

[6] Home — mbweb. https://yknzhu.wixsite.com/mbweb. (Accessed on 11/04/2021).

[7] News dataset available – common crawl. https://commoncrawl.org/2016/10/news-dataset-available/. (Accessed on 11/04/2021).

[8] News dataset available – common crawl. https://commoncrawl.org/2016/10/news-dataset-available/. (Accessed on 11/04/2021).

[9] Pubmed. https://pubmed.ncbi.nlm.nih.gov/. (Accessed on 10/28/2021).

[10] Semantic scholar — ai-powered research tool. https://www.semanticscholar.org/. (Accessed on 11/02/2021).

[11] Understanding lstm networks – colah's blog. https://colah.github.io/posts/2015-08-Understanding-LSTMs/. (Accessed on 10/31/2021).

[12] Using deepspeed and megatron to train megatron-turing nlg 530b, the world's largest and most powerful generative language model — nvidia developer blog. https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest (Accessed on 11/05/2021).

[13] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

[14] Xavier Amatriain. Transformer models: an introduction and catalog. *arXiv preprint arXiv:2302.07730*, 2023.

[15] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[16] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022.

[17] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[18] Haoping Bai, Meng Cao, Ping Huang, and Jiulong Shan. Self-supervised semi-supervised learning for data labeling and quality evaluation. *arXiv preprint arXiv:2111.10932*, 2021.

[19] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

[20] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[21] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[22] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando De Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. 2003.

[23] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. arxiv e-prints, art. *arXiv preprint arXiv:1206.5538*, 2012.

[24] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[25] Tamara L Berg, Alexander C Berg, Jaety Edwards, Michael Maire, Ryan White, Yee-Whye Teh, Erik Learned-Miller, and David A Forsyth. Names and faces in the news. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004.

[26] David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, 2003.

[27] Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019.

[28] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[29] Mikhail S Burtsev, Yuri Kuratov, Anton Peganov, and Grigory V Sapunov. Memory transformer. *arXiv preprint arXiv:2006.11527*, 2020.

[30] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[31] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[32] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[33] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

[34] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12655–12663, 2020.

[35] Minghao Chen, Houwen Peng, Jianlong Fu, and Haibin Ling. Autoformer: Searching transformers for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12270–12280, 2021.

[36] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022.

[37] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[38] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[39] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022.

[40] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[41] Yushi Chen, Hanlu Jiang, Chunyang Li, Xiuping Jia, and Pedram Ghamisi. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 54(10):6232–6251, 2016.

[42] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[43] Wang Chong, David Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *2009 IEEE Conference on computer vision and pattern recognition*, pages 1903–1910. IEEE, 2009.

[44] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *arXiv preprint arXiv:1506.07503*, 2015.

[45] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[46] Lee Ad Cooper, Elizabeth G Demicco, Joel H Saltz, Reid T Powell, Arvind Rao, and Alexander J Lazar. Pancancer insights from the cancer genome atlas: the pathologist's perspective. *The Journal of pathology*, 244(5):512–524, 2018.

[47] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. Multi-head attention: Collaborate instead of concatenate. *arXiv preprint arXiv:2006.16362*, 2020.

[48] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. *Advances in neural information processing systems*, 28:3079–3087, 2015.

[49] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[50] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[51] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. Technical report, Technical Report, 2021.

[52] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[53] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv preprint arXiv:2107.00652*, 2021.

[54] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[55] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021.

[56] Maha Elbayad, Laurent Besacier, and Jakob Verbeek. Pervasive attention: 2d convolutional neural networks for sequence-to-sequence prediction. *arXiv preprint arXiv:1808.03867*, 2018.

[57] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[58] A Frome, GS Corrado, J Shlens, et al. A deep visual-semantic embedding model. *Proceedings of the Advances in Neural Information Processing Systems*, pages 2121–2129.

[59] Jevgenij Gamper and Nasir Rajpoot. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16549–16559, 2021.

[60] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[61] Tobias Glasmachers. Limits of end-to-end learning. In *Asian conference on machine learning*, pages 17–32. PMLR, 2017.

[62] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[63] Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):855–868, 2008.

[64] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[65] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7181–7189, 2018.

[66] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*, 2020.

[67] Trevor Hastie, Robert Tibshirani, Jerome Friedman, Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Overview of supervised learning. *The elements of statistical learning: Data mining, inference, and prediction*, pages 9–41, 2009.

[68] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[70] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[71] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[72] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2310–2318, 2017.

[73] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018.

[74] Marie Humbert-Droz, Pritam Mukherjee, Olivier Gevaert, et al. Strategies to address the lack of labeled data for supervised machine learning training with electronic health records: case study for the extraction of symptoms from clinical notes. *JMIR Medical Informatics*, 10(3):e32903, 2022.

[75] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, and Shanay Rab. Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, 3:58–73, 2022.

[76] Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. Saliency-guided attention network for image-sentence matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5754–5763, 2019.

[77] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

[78] Shivam Kalra, Mohammed Adnan, Graham Taylor, and Hamid R Tizhoosh. Learning permutation invariant representations using memory networks. In *European Conference on Computer Vision*, pages 677–693. Springer, 2020.

[79] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[80] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *arXiv preprint arXiv:1406.5679*, 2014.

[81] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[82] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021.

[83] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[84] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[85] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[86] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. *arXiv preprint arXiv:2102.04432*, 2021.

[87] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[88] Van-Duc Le. Veml: An end-to-end machine learning lifecycle for large-scale and high-dimensional data. *arXiv preprint arXiv:2304.13037*, 2023.

[89] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[90] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4654–4662, 2019.

[91] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[92] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 2022.

[93] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[94] Jing Liu, Changsheng Xu, and Hanqing Lu. Cross-media retrieval: state-of-the-art and open issues. *International Journal of Multimedia Intelligence and Security*, 1(1):33–52, 2010.

[95] Lu Liu, William Hamilton, Guodong Long, Jing Jiang, and Hugo Larochelle. A universal representation transformer layer for few-shot image classification. *arXiv preprint arXiv:2006.11702*, 2020.

[96] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[97] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[98] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[99] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[100] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[101] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Andrew Zhang, Long Phi Le, et al. Towards a visual-language foundation model for computational pathology. *arXiv preprint arXiv:2307.12914*, 2023.

[102] Xiaopeng Lu, Tiancheng Zhao, and Kyusong Lee. Visualsparta: An embarrassingly simple approach to large-scale text-to-image search with weighted bag-of-words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5020–5029, 2021.

[103] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[104] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision*, pages 2623–2631, 2015.

[105] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical image analysis*, 33:170–175, 2016.

[106] Danial Maleki and Hamid R Tizhoosh. Lile: Look in-depth before looking elsewhere–a dual attention network using transformers for cross-modal information retrieval in histopathology archives. In *International Conference on Medical Imaging with Deep Learning*, pages 879–894. PMLR, 2022.

[107] Shahid Mehmood, Taher M Ghazal, Muhammad Adnan Khan, Muhammad Zubair, Muhammad Tahir Naseem, Tauqeer Faiz, and Munir Ahmad. Malignancy detection in lung and colon histopathology images using transfer learning with class selective image processing. *IEEE Access*, 10:25657–25668, 2022.

[108] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *arXiv preprint arXiv:2105.10497*, 2021.

[109] Mark Nixon and Alberto Aguado. *Feature extraction and image processing for computer vision*. Academic press, 2019.

[110] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[111] OpenAI. Gpt-4 technical report. *arXiv*, 2023.

[112] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby,

et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[113] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020.

[114] Yuxin Peng and Jinwei Qi. Cm-gans: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1):1–24, 2019.

[115] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[116] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[117] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[118] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

[119] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[120] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[121] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[122] Abtin Riasatian, Morteza Babaie, Danial Maleki, Shivam Kalra, Mojtaba Valipour, Sobhan Hemati, Manit Zaveri, Amir Safarpoor, Sobhan Shafiei, Mehdi Afshari, et al. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Medical Image Analysis*, 70:102032, 2021.

[123] Salva Rühling Cachay, Benedikt Boecking, and Artur Dubrawski. End-to-end weak supervision. *Advances in Neural Information Processing Systems*, 34:1845–1857, 2021.

[124] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[125] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[126] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[127] Danijel Skocaj, Ales Leonardis, and Geert-Jan M Kruijff. Cross-modal learning. *Encyclopedia of the Sciences of Learning*, pages 861–864, 2012.

[128] Richard Socher and Li Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 966–973. IEEE, 2010.

[129] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[130] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pages 1857–1865, 2016.

[131] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019.

[132] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[133] Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *ICML*, 2011.

[134] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[135] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

[136] Christopher Thomas and Adriana Kovashka. Preserving semantic neighborhoods for robust cross-modal retrieval. In *European Conference on Computer Vision*, pages 317–335. Springer, 2020.

[137] Christopher Thomas and Adriana Kovashka. Emphasizing complementary samples for non-literal cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4632–4641, 2022.

[138] Mesut Toğaçar. Disease type detection in lung and colon cancer images using the complement approach of inefficient sets. *Computers in Biology and Medicine*, 137:104827, 2021.

[139] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[140] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. *arXiv preprint arXiv:2103.17239*, 2021.

[141] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[142] Masayuki Tsuneki and Fahdi Kanavati. Inference of captions from histopathological patches. In *International Conference on Medical Imaging with Deep Learning*, pages 1235–1250. PMLR, 2022.

[143] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[144] Alex Jinpeng Wang, Pan Zhou, Mike Zheng Shou, and Shuicheng Yan. Position-guided text prompt for vision-language pre-training. *arXiv preprint arXiv:2212.09737*, 2022.

[145] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1508–1517, 2020.

[146] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[147] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. Sceneformer: Indoor scene generation with transformers. *arXiv preprint arXiv:2012.09793*, 2020.

[148] Yanan Wang, Litao Yang, Geoffrey I Webb, Zongyuan Ge, and Jiangning Song. Octid: a one-class learning-based python package for tumor image detection. *Bioinformatics*, 37(21):3986–3988, 2021.

[149] Yiding Wang, Decang Sun, Kai Chen, Fan Lai, and Mosharaf Chowdhury. Efficient dnn training with knowledge-guided layer freezing. *arXiv preprint arXiv:2201.06227*, 2022.

[150] Yu-An Wang and Yun-Nung Chen. What do position embeddings learn? an empirical study of pre-trained language model positional encoding. *arXiv preprint arXiv:2010.04903*, 2020.

[151] Lilian Weng. Attention? attention! *lilianweng.github.io/lil-log*, 2018.

[152] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[153] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng. P2t: Pyramid pooling transformer for scene understanding. *arXiv preprint arXiv:2106.12011*, 2021.

[154] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[155] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020.

[156] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10502–10511, 2019.

[157] W Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[158] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[159] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[160] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[161] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.

[162] Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. X ^{2}-vlm: All-in-one pre-trained model for vision-language tasks. *arXiv preprint arXiv:2211.12402*, 2022.

[163] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021.

[164] Hongjie Zhang. Unified framework for feature extraction based on contrastive learning. *arXiv preprint arXiv:2101.11703*, 2021.

[165] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023.

[166] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657, 2015.

[167] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436, 2017.

[168] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020.

[169] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[170] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.